

Trace Pursuit: A General Framework for Model-Free Variable Selection

Zhou Yu^{*}, Yuexiao Dong[†], and Li-Xing Zhu[‡]

^{*}East China Normal University,

[†]Temple University, and

[‡]The Hong Kong Baptist University

[‡]*Address for correspondence:* Li-Xing Zhu, Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong. Email: lzhu@hkbu.edu.hk.

Trace Pursuit: A General Framework for Model-Free Variable Selection

Abstract

We propose trace pursuit for model-free variable selection under the sufficient dimension reduction paradigm. Two distinct algorithms are proposed: stepwise trace pursuit and forward trace pursuit. Stepwise trace pursuit achieves selection consistency with fixed p , and is readily applicable in the challenging setting with $p > n$. Forward trace pursuit can serve as an initial screening step to speed up the computation in the case of ultrahigh dimensionality. The screening consistency property of forward trace pursuit based on sliced inverse regression is established. Finite sample performances of trace pursuit and other model-free variable selection methods are compared through numerical studies.

Key Words: directional regression, sliced average variance estimation, selection consistency, sliced inverse regression, stepwise regression.

1. Introduction

Contemporary statistical analysis often encounters high dimensional datasets that are routinely collected in a wide range of research areas, where the predictor dimensionality may easily dominate the relatively small sample size. To include the significant variables and exclude the insignificant variables at the same time, the paradigm of variable selection has seen much progress in recent years. Many popular variable selection procedures are developed under the linear model or the generalized linear model framework, such as nonnegative garrotte (Breiman, 1995), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), group LASSO (Yuan and Lin, 2006), Dantzig selector (Candés and Tao, 2007), and MCP (Zhang, 2010).

Let $\mathbf{X} = (x_1, \dots, x_p)^T$ be the predictor and Y be the scalar response. Model-free variable selection aims to find the index set \mathcal{A} such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}, \quad (1.1)$$

where $\perp\!\!\!\perp$ stands for independence, \mathcal{A}^c is the complement of \mathcal{A} in the index set $\mathcal{I} = \{1, \dots, p\}$, $\mathbf{X}_{\mathcal{A}} = \{x_i : i \in \mathcal{A}\}$, and $\mathbf{X}_{\mathcal{A}^c} = \{x_i : i \in \mathcal{A}^c\}$. Condition (1.1) implies that $\mathbf{X}_{\mathcal{A}}$ contains all the active predictors in terms of predicting Y . Ideally, we want to find the smallest index set \mathcal{A} satisfying (1.1), in which case no inactive predictors are included in $\mathbf{X}_{\mathcal{A}}$. Model-free variable selection is closely related to sufficient dimension reduction (Li, 1991; Cook, 1998), which aims to find subspace \mathcal{S} such that

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}} \mathbf{X}. \quad (1.2)$$

Here $P_{(\cdot)}$ denotes the projection operator with respect to the standard inner product.

Under mild conditions (Yin et al., 2008), the intersection of all \mathcal{S} satisfying (1.2) still satisfies (1.2). We call this intersection the central space and denote it by $\mathcal{S}_{Y|\mathbf{X}}$. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is called the structural dimension and we denote it by q with $q < p$. Some popular sufficient dimension reduction methods in the literature include sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), and directional regression (Li and Wang, 2007).

There are two distinct approaches in the literature for model-free variable selection: the sparse sufficient dimension reduction approach and the hypothesis testing approach. By noting that many dimension reduction methods could be reformulated as a least squares problem, Li (2007) proposed sparse sufficient dimension reduction by combining sufficient dimension reduction with penalized least squares. Other sparse dimension reduction methods include shrinkage SIR (Ni et al., 2005), constrained canonical correlation (Zhou and He, 2008), and regularized SIR (Li and Yin, 2008). While traditional sufficient dimension reduction finds linear combinations of all the original variables, sparse sufficient dimension reduction achieves dimension reduction and variable selection simultaneously. The state of the art method in this category is the coordinate-independent sparse dimension reduction (CISE) (Chen et al., 2010), where a subspace-oriented penalty is proposed such that the resulting central space has the same sparsity structure regardless of the chosen basis of $\mathcal{S}_{Y|\mathbf{X}}$. Although it enjoys the oracle property that it performs asymptotically as well as if the true irrelevant predictors were known, CISE is not applicable when p is larger than the sample size n .

Model-free variable selection through sufficient dimension reduction can also be implemented under the hypothesis testing framework. Without loss of generality, we assume the active index set $\mathcal{A} = \{1, \dots, K\}$ for ease of demonstration. Then (1.1) is

equivalent to $P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$, where $\mathcal{H} = \text{Span} \{(\mathbf{0}_{(p-K) \times K}, \mathbf{I}_{p-K})^T\}$ is the subspace of \mathbb{R}^p corresponding to the coordinates of the inactive predictors, and \mathcal{O}_p denotes the origin in \mathbb{R}^p . To test $H_0 : P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$ versus $H_a : P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} \neq \mathcal{O}_p$, Cook (2004) proposed the marginal coordinate hypothesis test based on SIR, and a similar test based on SAVE was developed in Shao et al. (2007). Backward elimination for variable selection based on such tests is discussed in Li et al. (2005). However, these tests rely on an initial estimator of the central space $\mathcal{S}_{Y|\mathbf{X}}$ via SIR or SAVE, which is not available when $p > n$.

To achieve model-free variable selection with $p > n$, Zhong et al. (2012) proposed correlation pursuit (COP). COP looks for a subset of variables in \mathbf{X} to maximize an objective function, which measures the correlation between the transformed response Y and the projections of \mathbf{X} . COP is based on SIR and inherits the limitations of SIR. Namely, COP may miss significant predictors linked to the response through quadratic functions or interactions. More recently, Jiang and Liu (2013) proposed a likelihood ratio test based procedure named SIR with interaction detection (SIRI). SIRI includes a special case that is asymptotically equivalent to COP, and it extends COP by detecting significant predictors that appear in interactions. Both COP and SIRI involve estimation of the structural dimension q of $\mathcal{S}_{Y|\mathbf{X}}$, which is known as order determination in the sufficient dimension reduction literature. Order determination in the $p > n$ setting is a challenging issue, and the performances of COP and SIRI may deteriorate when the structural dimension q can not be accurately estimated.

We propose trace pursuit as a novel approach for model-free variable selection in this paper. Based on the newly designed method-specific (SIR, SAVE, or directional regression) trace tests, we first extend the classical stepwise regression in linear models and propose a stepwise trace pursuit (STP) algorithm for model-free variable selection.

STP iterates between adding one predictor from outside the working index set \mathcal{F} and deleting one predictor from within \mathcal{F} . Furthermore, we mimic the forward regression in the linear model setting and propose the forward trace pursuit (FTP) algorithm. After finding a solution path by adding one predictor into the model at a time, a modified BIC criterion provides a chosen model that is guaranteed to include all important predictors. Finally, our two-stage hybrid trace pursuit (HTP) algorithm uses FTP for initial variable screening, which is followed by STP for the refined variable selection at the second stage. While SIR-based HTP might miss some significant predictors involved only in interactions and SAVE-based HTP may miss significant predictors that is linked to the response through a linear function, HTP based on directional regression can successfully detect predictors in a wide range of models. Compared with existing methods in the literature, the trace pursuit: (1) can be combined with different existing sufficient dimension reduction methods to detect significant predictors linked through various unknown functions to the response; (2) does not rely on estimation of the structural dimension q ; (3) is designed to deal with the challenging $p > n$ setting; (4) provides a unified framework for model-free variable screening through FTP and model-free variable selection through STP. The selection consistency of the STP algorithms as well as the screening consistency property of the SIR-based FTP algorithm are established.

The paper is organized as follows. We first propose the SIR-based trace test and then extend it to SAVE-based and directional regression-based trace tests in Section 2. The asymptotic distributions of the proposed test statistics are discussed in Section 3. The STP algorithm and its selection consistency property are developed in Section 4. FTP for screening and HTP for two-stage model-free variable selection are discussed in Section 5. Section 6 provides some numerical studies including a real data analysis. Section 7 concludes the paper with some discussions.

2. Principle of the trace test

2.1. Some preliminaries

We briefly review three popular sufficient dimension reduction methods, SIR, SAVE and directional regression. Without loss of generality, assume $E(\mathbf{X}) = \mathbf{0}$ and $E(Y) = 0$. Let $\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$ and $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ denotes the standardized predictor. Suppose $\boldsymbol{\beta} \in \mathbb{R}^{p \times q}$ is the basis of $\mathcal{S}_{Y|\mathbf{X}}$ and $\boldsymbol{\eta} \in \mathbb{R}^{p \times q}$ is the basis of the \mathbf{Z} -scaled central space $\mathcal{S}_{Y|\mathbf{Z}}$. Let $\{J_1, \dots, J_H\}$ be a measurable partition of Ω_Y , the sample space of Y . The kernel matrix of the classical SIR (Li, 1991) is defined as $\mathbf{M}^{\text{SIR}} = \text{Var} \{E(\mathbf{Z}|Y \in J_h)\}$. Under the linear conditional mean (LCM) assumption that

$$E(\mathbf{X}|\boldsymbol{\beta}^T \mathbf{X}) \text{ is a linear function of } \boldsymbol{\beta}^T \mathbf{X}, \quad (2.1)$$

we have $\text{Span}(\mathbf{M}^{\text{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Here $\text{Span}(\mathbf{M})$ denotes the column space of \mathbf{M} . Under the additional constant conditional variance (CCV) assumption that

$$\text{Var}(\mathbf{X}|\boldsymbol{\beta}^T \mathbf{X}) \text{ is nonrandom}, \quad (2.2)$$

Cook and Weisberg (1991) demonstrate that $\text{Span}(\mathbf{M}^{\text{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$, where $\mathbf{M}^{\text{SAVE}} = E \{\mathbf{I}_p - \text{Var}(\mathbf{Z}|Y \in J_h)\}^2$ is the kernel matrix for SAVE. When \mathbf{X} is normal, both LCM and CCV assumptions are satisfied. For nonnormal predictor \mathbf{X} , please refer to Cook and Nachtsheim (1994), Cook and Li (2009), Li and Dong (2009), Dong and Li (2010).

It is well-known that SIR and SAVE are complement to each other in both the regression and the classification settings. SIR works better when the link function between the continuous response and the predictor is monotone, or when there is location shift

between different categories of the discrete response. SAVE, on the other hand, is more effective with U-shaped link function or detecting scale difference. Directional regression is designed to combine the strength of SIR and SAVE. For kernel matrix

$$\begin{aligned}\mathbf{M}^{\text{DR}} = & 2E\{E^2(\mathbf{Z}\mathbf{Z}^T|Y \in J_h)\} + 2E^2\{E(\mathbf{Z}|Y \in J_h)E^T(\mathbf{Z}|Y \in J_h)\} \\ & + 2E\{E^T(\mathbf{Z}|Y \in J_h)E(\mathbf{Z}|Y \in J_h)\}E\{E(\mathbf{Z}|Y \in J_h)E^T(\mathbf{Z}|Y \in J_h)\} - 2\mathbf{I}_p,\end{aligned}$$

Li and Wang (2007) prove that $\text{Span}(\mathbf{M}^{\text{DR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ under assumptions (2.1) and (2.2).

2.2. SIR-based trace test

We state the principle of the SIR-based trace test in this section. For working index set \mathcal{F} and index $j \in \mathcal{F}^c$, we want to test

$$H_0 : Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}} \text{ v.s. } H_a : Y \text{ is not independent of } x_j \text{ given } \mathbf{X}_{\mathcal{F}}. \quad (2.3)$$

Denote $R_h = I(Y \in J_h)$, $p_h = E(R_h)$, and $\mathbf{U}_h = E(\mathbf{X}|Y \in J_h)$. The kernel matrix for SIR can be rewritten as $\mathbf{M}^{\text{SIR}} = \Sigma^{-1/2} \left(\sum_{h=1}^H p_h \mathbf{U}_h \mathbf{U}_h^T \right) \Sigma^{-1/2}$. For any index set \mathcal{F} , denote $\mathbf{X}_{\mathcal{F}} = \{x_i : i \in \mathcal{F}\}$, $\text{Var}(\mathbf{X}_{\mathcal{F}}) = \Sigma_{\mathcal{F}}$, and $\mathbf{U}_{\mathcal{F},h} = E(\mathbf{X}_{\mathcal{F}}|Y \in J_h)$. We mimic \mathbf{M}^{SIR} and define $\mathbf{M}_{\mathcal{F}}^{\text{SIR}}$ as

$$\mathbf{M}_{\mathcal{F}}^{\text{SIR}} = \Sigma_{\mathcal{F}}^{-1/2} \left(\sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T \right) \Sigma_{\mathcal{F}}^{-1/2}. \quad (2.4)$$

Recall that \mathcal{A} denotes the active index set satisfying $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$, and $\mathcal{I} = \{1, \dots, p\}$ denotes the full index set. We have the following key observation.

Proposition 2.1. *Suppose the LCM assumption (2.1) holds true. Then for any index*

set \mathcal{F} such that $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, we have $\text{tr}(\mathbf{M}_{\mathcal{A}}^{\text{SIR}}) = \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = \text{tr}(\mathbf{M}^{\text{SIR}})$.

Proposition 2.1 suggests that we use $\text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})$ to capture the strength of relationship between Y and $\mathbf{X}_{\mathcal{F}}$. Denote $\mathcal{F} \cup j$ as the index set of j together with all the indices in \mathcal{F} . Given that $\mathbf{X}_{\mathcal{F}}$ is already in the model, we will see that the trace difference $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})$ can be used to test the contribution of the additional variable x_j to Y . The following subset LCM assumption is required before we state the main result,

$$E(x_j|\mathbf{X}_{\mathcal{F}}) \text{ is a linear function of } \mathbf{X}_{\mathcal{F}} \text{ for any } \mathcal{F} \subset \mathcal{I} \text{ and } j \in \mathcal{F}^c. \quad (2.5)$$

Assumption (2.5) is parallel to the LCM assumption (2.1), and both are satisfied when \mathbf{X} is elliptically contour distributed. The principle of the SIR-based trace test is stated in the next theorem.

Theorem 2.1. *Assume the subset LCM assumption (2.5) holds true. Then for $\mathcal{F} \subset \mathcal{I}$ and $j \in \mathcal{F}^c$, we have*

1. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = \sum_{h=1}^H p_h \gamma_{j|\mathcal{F},h}^2$, where $\gamma_{j|\mathcal{F},h} = E(\gamma_{j|\mathcal{F}}|Y \in J_h)$ with $x_{j|\mathcal{F}} = x_j - E(x_j|\mathbf{X}_{\mathcal{F}})$, $\sigma_{j|\mathcal{F}}^2 = \text{Var}(x_{j|\mathcal{F}})$, and $\gamma_{j|\mathcal{F}} = x_{j|\mathcal{F}}/\sigma_{j|\mathcal{F}}$.
2. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = 0$ given that $\mathcal{A} \subseteq \mathcal{F}$.

Part 1 of Theorem 2.1 provides the explicit formula to calculate the trace difference between $\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}$ and $\mathbf{M}_{\mathcal{F}}^{\text{SIR}}$. Part 2 of Theorem 2.1 states that the trace difference is zero when the working index set \mathcal{F} contains the active set \mathcal{A} .

The idea of using trace difference is similar to the extra sums of squares test in the classical multiple linear regression setting. Zhong et al. (2012) suggest a related test in the COP algorithm. Given the structural dimension q , denote the largest q eigenvalues

of $\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}$ as $\lambda_{\mathcal{F} \cup j}^{(k)}$, and the largest q eigenvalues of $\mathbf{M}_{\mathcal{F}}^{\text{SIR}}$ as $\lambda_{\mathcal{F}}^{(k)}$, $k = 1, \dots, q$. COP is based on the key quantity $\sum_{k=1}^q (1 - \lambda_{\mathcal{F} \cup j}^{(k)})^{-1} (\lambda_{\mathcal{F} \cup j}^{(k)} - \lambda_{\mathcal{F}}^{(k)})$. The COP test reduces to the trace test with $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})$ if we drop the scaling factor $(1 - \lambda_{\mathcal{F} \cup j}^{(k)})^{-1}$ and assume both $\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}$ and $\mathbf{M}_{\mathcal{F}}^{\text{SIR}}$ have rank q . Compared with the COP test, the SIR-based trace test does not involve estimating q . While COP can not be easily extended to dimension reduction methods other than SIR, the trace test is a versatile framework and can be combined with methods other than SIR, as we will see next.

2.3. SAVE-based and directional regression-based trace tests

Note that \mathbf{M}^{SAVE} can be rewritten as $\sum_{h=1}^H p_h \{\Sigma^{-1/2}(\Sigma - \mathbf{V}_h + \mathbf{U}_h \mathbf{U}_h^T) \Sigma^{-1/2}\}^2$ with $\mathbf{V}_h = E(\mathbf{X} \mathbf{X}^T | Y \in J_h)$. Denote $\mathbf{V}_{\mathcal{F}, h} = E(\mathbf{X}_{\mathcal{F}} \mathbf{X}_{\mathcal{F}}^T | Y \in J_h)$ and define

$$\mathbf{M}_{\mathcal{F}}^{\text{SAVE}} = \sum_{h=1}^H p_h \{\Sigma_{\mathcal{F}}^{-1/2}(\Sigma_{\mathcal{F}} - \mathbf{V}_{\mathcal{F}, h} + \mathbf{U}_{\mathcal{F}, h} \mathbf{U}_{\mathcal{F}, h}^T) \Sigma_{\mathcal{F}}^{-1/2}\}^2.$$

For the purpose of sufficient dimension reduction, it is well-known that SAVE requires the CCV assumption (2.2) in addition to the LCM assumption (2.1) required by SIR. In a parallel fashion, our SAVE-based trace test relies on the following subset CCV assumption together with the subset LCM assumption (2.5),

$$\text{Var}(x_j | \mathbf{X}_{\mathcal{F}}) \text{ is nonrandom for any } \mathcal{F} \subset \mathcal{I} \text{ and } j \in \mathcal{F}^c. \quad (2.6)$$

Assumptions (2.1) and (2.2) are common in the sufficient dimension reduction literature. Meanwhile, assumptions (2.5) and (2.6) have been used in Zhong et al. (2012) and Jiang and Liu (2013) for SIR based model-free variable selection. All four conditions are satisfied when \mathbf{X} is normal. The next theorem states the principle of the SAVE-based trace test.

Theorem 2.2. Assume the subset LCM assumption (2.5) and the subset CCV assumption (2.6) hold true. Then for $\mathcal{F} \subset \mathcal{I}$ and $j \in \mathcal{F}^c$, we have

1. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SAVE}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SAVE}}) = \sum_{h=1}^H p_h \{(1 - \zeta_{j|\mathcal{F},h} + \gamma_{j|\mathcal{F},h}^2)^2 + 2\phi_{j|\mathcal{F},h}^T \phi_{j|\mathcal{F},h}\}$, where $\phi_{j|\mathcal{F},h} = \Sigma_{\mathcal{F}}^{-1/2} \{\mathbf{U}_{\mathcal{F},h} \gamma_{j|\mathcal{F},h} - E(\mathbf{X}_{\mathcal{F}} \gamma_{j|\mathcal{F}} | Y \in J_h)\}$ and $\zeta_{j|\mathcal{F},h} = E(\gamma_{j|\mathcal{F}}^2 | Y \in J_h)$.
2. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SAVE}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SAVE}}) = 0$ given that $\mathcal{A} \subseteq \mathcal{F}$.

For directional regression-based trace test, define

$$\begin{aligned} \mathbf{M}_{\mathcal{F}}^{\text{DR}} = & 2 \sum_{h=1}^H p_h (\Sigma_{\mathcal{F}}^{-1/2} \mathbf{V}_{\mathcal{F},h} \Sigma_{\mathcal{F}}^{-1/2})^2 + 2 \left(\sum_{h=1}^H p_h \Sigma_{\mathcal{F}}^{-1/2} \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T \Sigma_{\mathcal{F}}^{-1/2} \right)^2 \\ & + 2 \left(\sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F},h}^T \Sigma_{\mathcal{F}}^{-1} \mathbf{U}_{\mathcal{F},h} \right) \left(\sum_{h=1}^H p_h \Sigma_{\mathcal{F}}^{-1/2} \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T \Sigma_{\mathcal{F}}^{-1/2} \right) - 2\mathbf{I}_{|\mathcal{F}|}, \end{aligned}$$

where $|\mathcal{F}|$ denotes the cardinality of \mathcal{F} . The directional regression-based trace test relies on the next theorem.

Theorem 2.3. Assume the subset LCM assumption (2.5) and the subset CCV assumption (2.6) hold true. Then for $\mathcal{F} \subset \mathcal{I}$ and $j \in \mathcal{F}^c$, we have

1. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{DR}}) = 2 \sum_{h=1}^H p_h \left((1 - \zeta_{j|\mathcal{F},h})^2 + 2\boldsymbol{\nu}_{j|\mathcal{F},h}^T \boldsymbol{\nu}_{j|\mathcal{F},h} \right) + 4\varrho_{j|\mathcal{F}}^2 + 4\boldsymbol{\iota}_{j|\mathcal{F}}^T \boldsymbol{\iota}_{j|\mathcal{F}} + 4\kappa_{\mathcal{F}} \varrho_{j|\mathcal{F}}$, where $\boldsymbol{\nu}_{j|\mathcal{F},h} = \Sigma_{\mathcal{F}}^{-1/2} E(\mathbf{X}_{\mathcal{F}} \gamma_{j|\mathcal{F}} | Y \in J_h)$, $\boldsymbol{\iota}_{j|\mathcal{F},h} = \Sigma_{\mathcal{F}}^{-1/2} \mathbf{U}_{\mathcal{F},h} \gamma_{j|\mathcal{F},h}$, $\boldsymbol{\iota}_{j|\mathcal{F}} = \sum_{h=1}^H p_h \boldsymbol{\iota}_{j|\mathcal{F},h}$, $\varrho_{j|\mathcal{F}} = \sum_{h=1}^H p_h \gamma_{j|\mathcal{F},h}^2$, and $\kappa_{\mathcal{F}} = \sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F},h}^T \Sigma_{\mathcal{F}}^{-1} \mathbf{U}_{\mathcal{F},h}$.
2. $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{DR}}) = 0$ given that $\mathcal{A} \subseteq \mathcal{F}$.

Theorems 2.2 and 2.3 demonstrate that the trace test can be a general framework. Unlike COP, trace tests do not require estimation of the structural dimension q , and they can be combined with sufficient dimension reduction methods other than SIR.

3. Asymptotic distributions of the trace test statistics

Given an i.i.d. sample (\mathbf{X}_i, Y_i) , $i = 1, \dots, n$, we develop the asymptotic distribution of the sample level trace test statistics. The asymptotic results in this section are developed with fixed $|\mathcal{F}|$ when n goes to infinity. For the SIR-based test, we have

Theorem 3.1. *Suppose \mathbf{X} has finite fourth order moment, and the subset LCM assumption (2.5) holds true. Then under $H_0 : Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$, $j \in \mathcal{F}^c$, we have*

$$T_{j|\mathcal{F}}^{\text{SIR}} \longrightarrow \sum_{k=1}^H \omega_{j|\mathcal{F},k}^{\text{SIR}} \chi_1^2, \text{ where } T_{j|\mathcal{F}}^{\text{SIR}} = n \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) \right\}.$$

Here $\omega_{j|\mathcal{F},1}^{\text{SIR}} \geq \dots \geq \omega_{j|\mathcal{F},H}^{\text{SIR}}$ are the eigenvalues of $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{SIR}}$ defined in the Appendix.

The test statistic $T_{j|\mathcal{F}}^{\text{SIR}}$ can be calculated as $n \sum_{h=1}^H \hat{p}_h \hat{\gamma}_{j|\mathcal{F},h}^2$, where \hat{p}_h and $\hat{\gamma}_{j|\mathcal{F},h}$ are sample counterparts of p_h and $\gamma_{j|\mathcal{F},h}$ defined in Theorem 2.1. Since we assume $E(\mathbf{X}) = \mathbf{0}$ for the population level development, $\hat{\gamma}_{j|\mathcal{F},h}$ is calculated based on centered predictors. Let $\tilde{\mathbf{X}}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{ip})^T = \mathbf{X}_i - \sum_{i=1}^n \mathbf{X}_i / n$ be the centered version of \mathbf{X}_i . Denote $R_{i,h} = I(Y_i \in J_h)$, $\hat{p}_h = \sum_{i=1}^n R_{i,h} / n$, $\tilde{\mathbf{X}}_{i(\mathcal{F})} = \{\tilde{x}_{ij} : j \in \mathcal{F}\}$, and $\hat{\mathbf{U}}_{\mathcal{F},h} = \sum_{i=1}^n \tilde{\mathbf{X}}_{i(\mathcal{F})} R_{i,h} / (n\hat{p}_h)$. Let $E_n(\tilde{x}_j | \tilde{\mathbf{X}}_{\mathcal{F}})$ be the sample estimator of $E(\tilde{x}_j | \tilde{\mathbf{X}}_{\mathcal{F}})$. Further denote $\hat{x}_{ij|\mathcal{F}} = \tilde{x}_{ij} - E_n(\tilde{x}_j | \tilde{\mathbf{X}}_{\mathcal{F}})$, $\hat{\sigma}_{j|\mathcal{F}}^2 = \sum_{i=1}^n \hat{x}_{ij|\mathcal{F}}^2 / n - (\sum_{i=1}^n \hat{x}_{ij|\mathcal{F}})^2 / n^2$, and $\hat{\gamma}_{ij|\mathcal{F}} = \hat{x}_{ij|\mathcal{F}} / \hat{\sigma}_{j|\mathcal{F}}$. Then we have $\hat{\gamma}_{j|\mathcal{F},h} = \sum_{i=1}^n \hat{\gamma}_{ij|\mathcal{F}} R_{i,h} / (n\hat{p}_h)$.

The next two Theorems provide the asymptotic distribution for the SAVE-based and directional regression-based trace test statistics respectively.

Theorem 3.2. *Suppose \mathbf{X} has finite fourth order moment. Assume the subset LCM assumption (2.5) and the subset CCV assumption (2.6) hold true. Then under $H_0 :$*

$Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}, j \in \mathcal{F}^c$, we have

$$T_{j|\mathcal{F}}^{\text{SAVE}} \longrightarrow \sum_{k=1}^{(|\mathcal{F}|+1)H} \omega_{j|\mathcal{F},k}^{\text{SAVE}} \chi_1^2, \text{ where } T_{j|\mathcal{F}}^{\text{SAVE}} = n \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{SAVE}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SAVE}}) \right\}.$$

Here $\omega_{j|\mathcal{F},1}^{\text{SAVE}} \geq \dots \geq \omega_{j|\mathcal{F},(|\mathcal{F}|+1)H}^{\text{SAVE}}$ are the eigenvalues of $\mathbf{\Omega}_{j|\mathcal{F}}^{\text{SAVE}}$ defined in the Appendix.

Theorem 3.3. Suppose \mathbf{X} has finite fourth order moment. Assume the subset LCM assumption (2.5) and the subset CCV assumption (2.6) hold true. Then under H_0 : $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}, j \in \mathcal{F}^c$, we have

$$T_{j|\mathcal{F}}^{\text{DR}} \longrightarrow \sum_{k=1}^{2|\mathcal{F}|(H+1)} \omega_{j|\mathcal{F},k}^{\text{DR}} \chi_1^2, \text{ where } T_{j|\mathcal{F}}^{\text{DR}} = n \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{DR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{DR}}) \right\}.$$

Here $\omega_{j|\mathcal{F},1}^{\text{DR}} \geq \dots \geq \omega_{j|\mathcal{F},2|\mathcal{F}|(H+1)}^{\text{DR}}$ are the eigenvalues of $\mathbf{\Omega}_{j|\mathcal{F}}^{\text{DR}}$ defined in the Appendix.

From Theorem 2.2, we know $T_{j|\mathcal{F}}^{\text{SAVE}}$ can be calculated as $n \sum_{h=1}^H \hat{p}_h \{ (1 - \hat{\zeta}_{j|\mathcal{F},h} + \hat{\gamma}_{j|\mathcal{F},h}^2)^2 + 2\hat{\phi}_{j|\mathcal{F},h}^T \hat{\phi}_{j|\mathcal{F},h} \}$. From Theorem 2.3, we know $T_{j|\mathcal{F}}^{\text{DR}}$ can be calculated as $4n\hat{\varrho}_{j|\mathcal{F}}^2 + 4n\hat{\mathbf{l}}_{j|\mathcal{F}}^T \hat{\mathbf{l}}_{j|\mathcal{F}} + 4n\hat{\kappa}_{\mathcal{F}} \hat{\varrho}_{j|\mathcal{F}} + 2n \sum_{h=1}^H \hat{p}_h \left((1 - \hat{\zeta}_{j|\mathcal{F},h})^2 + 2\hat{\nu}_{j|\mathcal{F},h}^T \hat{\nu}_{j|\mathcal{F},h} \right)$. To approximate the asymptotic distribution under H_0 , we use estimated weights $\hat{\omega}_{j|\mathcal{F},k}^{\text{SIR}}$, $\hat{\omega}_{j|\mathcal{F},k}^{\text{SAVE}}$ and $\hat{\omega}_{j|\mathcal{F},k}^{\text{DR}}$. The detailed forms of these sample estimators are provided in the Appendix.

4. The stepwise trace pursuit algorithm

We provide the stepwise trace pursuit (STP) algorithm and its selection consistency property in this section. For the ease of presentation, the following stepwise algorithm is based on the SIR-based trace test. The STP algorithms for SAVE and directional regression can be defined in a parallel fashion.

- (a) *Initialization.* Set the initial working set to be $\mathcal{F} = \emptyset$.

(b) *Forward addition.* Find index $a_{\mathcal{F}}$ such that

$$a_{\mathcal{F}} = \arg \max_{j \in \mathcal{F}^c} \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{SIR}}). \quad (4.1)$$

If $T_{a_{\mathcal{F}}|\mathcal{F}}^{\text{SIR}} = n\{\text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup a_{\mathcal{F}}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}})\} > \bar{c}^{\text{SIR}}$, update \mathcal{F} to be $\mathcal{F} \cup a_{\mathcal{F}}$.

(c) *Backward deletion.* Find index $d_{\mathcal{F}}$ such that

$$d_{\mathcal{F}} = \arg \max_{j \in \mathcal{F}} \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \setminus j}^{\text{SIR}}). \quad (4.2)$$

If $T_{d_{\mathcal{F}}|\{\mathcal{F} \setminus d_{\mathcal{F}}\}}^{\text{SIR}} = n\{\text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \setminus d_{\mathcal{F}}}^{\text{SIR}})\} < \underline{c}^{\text{SIR}}$, update \mathcal{F} to be $\mathcal{F} \setminus d_{\mathcal{F}}$.

(d) Repeat steps (b) and (c) until no predictors can be added or deleted.

We provide some additional insight about the key quantity $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})$ before we study the selection consistency property of the SIR-based STP algorithm. Recall that \mathcal{A} is the true set of significant predictors, q denotes the structural dimension of the central space $\mathcal{S}_{Y|\mathbf{X}}$, and $\mathbf{M}^{\text{SIR}} = \text{Var}\{E(\mathbf{Z}|Y \in J_h)\}$ has q nonzero eigenvalues $\lambda_1 \geq \dots \geq \lambda_q$ with $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_q$ as the corresponding eigenvectors. Denote $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}_i = (\beta_{i,1}, \dots, \beta_{i,p})^T$ for $i = 1, \dots, q$, and let $\boldsymbol{\beta}_{\min} = \min_{j \in \mathcal{A}} \sqrt{\sum_{i=1}^q \beta_{i,j}^2}$.

Proposition 4.1. *Assume $\text{Span}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q) = \mathcal{S}_{Y|\mathbf{X}}$ and the subset LCM assumption (2.5) holds true. Then for any \mathcal{F} such that $\mathcal{F}^c \cap \mathcal{A} \neq \emptyset$, we have*

$$\max_{j \in \mathcal{F}^c \cap \mathcal{A}} \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} \geq \lambda_q \lambda_{\max}^{-1}(\boldsymbol{\Sigma}) \lambda_{\min}^2(\boldsymbol{\Sigma}) \boldsymbol{\beta}_{\min}^2,$$

where $\lambda_{\max}(\boldsymbol{\Sigma})$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ are the largest and the smallest eigenvalues of $\boldsymbol{\Sigma}$ respectively.

We have seen in Theorem 2.1 that $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = 0$ when $\mathcal{A} \subseteq \mathcal{F}$. Proposi-

tion 4.1 implies that when \mathcal{A} does not belong to \mathcal{F} , the maximum of $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})$ over $j \in \mathcal{F}^c \cap \mathcal{A}$ is greater than 0. In the linear regression setting, β_{\min} can be explained as the minimum signal strength, and it is common to assume that β_{\min} does not decrease to 0 too fast when n goes to infinity. This motivates us to assume that there exist $\varsigma > 0$ and $0 < \xi_{\min} < 1/2$ such that

$$\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} > \varsigma n^{-\xi_{\min}}. \quad (4.3)$$

Theorem 4.1. *Suppose \mathbf{X} has finite fourth order moment, condition (2.5) and condition (4.3) hold true.*

1. *If we set $0 < \bar{c}^{\text{SIR}} < \varsigma n^{1-\xi_{\min}}/2$, then as $n \rightarrow \infty$,*

$$Pr\left(\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} T_{j|\mathcal{F}}^{\text{SIR}} > \bar{c}^{\text{SIR}}\right) \rightarrow 1.$$

2. *If we set $\underline{c}^{\text{SIR}} > C n^{1-\xi_{\min}}$ for any $C > 0$, then as $n \rightarrow \infty$,*

$$Pr\left(\max_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} = \emptyset} \min_{j \in \mathcal{F}} T_{j|\{\mathcal{F} \setminus j\}}^{\text{SIR}} < \underline{c}^{\text{SIR}}\right) \rightarrow 1.$$

Part 1 of Theorem 4.1 implies that the addition step will not stop until all significant predictors are selected. Part 2 implies that the deletion step of the algorithm will not stop if the current selection includes any insignificant relevant predictors. Together, they guarantee the selection consistency of the STP algorithm for SIR. To guarantee the selection consistency of the STP algorithms for SAVE and the directional regression, condition (4.3) in Theorem 4.1 has to be updated accordingly. We leave the details to the Appendix. Note that the STP algorithm is directly applicable even with $p > n$. All

we need is that $|\mathcal{F}| < n$ for all iterations of the algorithm.

Condition (4.3) relates closely to the concept of exhaustiveness in the literature of sufficient dimension reduction. To fix the idea, consider a toy example $Y = x_1^2$ and x_i is i.i.d. $N(0, 1)$ for $i = 1, \dots, p$. It's easy to see that $\text{tr}(\mathbf{M}_{\mathcal{F} \cup \{1\}}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = 0$ for any \mathcal{F} , and condition (4.3) is violated. We expect the SIR-based STP algorithm to underfit when condition (4.3) is not satisfied.

5. The forward trace pursuit algorithm

To determine the index $a_{\mathcal{F}}$ in the addition step of the STP algorithm, we need to go over all possible candidate indices in \mathcal{F}^c and compare a total of $p - |\mathcal{F}|$ test statistics, which may lead to overwhelming computation burden when p is large. This motivates us to consider the forward trace pursuit (FTP) algorithm as an initial screening step. The SIR-based FTP algorithm is described as follows.

- (a) *Initialization.* Set $\mathcal{S}^{(0)} = \emptyset$.
- (b) *Forward addition.* For $k \geq 1$, $\mathcal{S}^{(k-1)}$ is given at the beginning of the k th iteration. For every $j \in \mathcal{I} \setminus \mathcal{S}^{(k-1)}$, compute $\text{tr}(\hat{\mathbf{M}}_{\mathcal{S}^{(k-1)} \cup j}^{\text{SIR}})$, and find a_k such that

$$a_k = \arg \max_{j \in \mathcal{I} \setminus \mathcal{S}^{(k-1)}} \text{tr}(\hat{\mathbf{M}}_{\mathcal{S}^{(k-1)} \cup j}^{\text{SIR}}).$$

- (c) *Solution path.* Repeat step (b) n times, to get a sequence of n nested candidate models. Denote the solution path as $\mathcal{S} = \{\mathcal{S}^{(k)} : 1 \leq k \leq n\}$, where $\mathcal{S}^{(k)} = \{a_1, \dots, a_k\}$.

To study the theoretical property of forward trace pursuit based on SIR, we assume the following set of conditions.

(C1) Assume that the predictor \mathbf{X} is normally distributed.

(C2) Assume that there exist two positive constant τ_{\min} and τ_{\max} , such that $\tau_{\min} <$

$$\lambda_{\min}(\boldsymbol{\Sigma}) < \lambda_{\max}(\boldsymbol{\Sigma}) < \tau_{\max} < \infty.$$

(C3) Assume condition (4.3) holds true, and there exist constants ξ and ξ_0 , such that

$$\log p \leq \varpi n^\xi, |\mathcal{A}| \leq \varpi n^{\xi_0}, \text{ and } \xi + 2\xi_{\min} + 2\xi_0 < 1.$$

(C1) and (C2) are commonly used conditions in high dimensional sparse covariance estimation and variable screening problems. Wang (2009) assumed (C1) and (C2) to study the sure screening property of forward linear regression. Condition (C3) allows both the predictor dimensionality p and the number of significant predictors $|\mathcal{A}|$ go to infinity as n goes to infinity. Note that (C1) implies condition (2.5). Denote $[t]$ as the smallest integer no less than t . We state the screening consistency of the SIR-based FTP algorithm next.

Theorem 5.1. *Assume conditions (C1)-(C3) hold true. Then as $n \rightarrow \infty$ and $p \rightarrow \infty$, the solution path of the SIR-based FTP algorithm satisfies*

$$Pr\left(\mathcal{A} \subset \mathcal{S}^{([2H\varsigma^{-1}\varpi n^{\xi_0+\xi_{\min}}])}\right) \rightarrow 1.$$

Theorem 5.1 guarantees that the FTP based on SIR enjoys the sure screening property in a model free setting, which extends the theoretical developments in Wang (2009). Moreover, Theorem 5.1 implies that with n going to infinity, only a finite number of iterations is needed in the FTP algorithm to recover the set \mathcal{A} of true significant predictors if the dimension of the true model is finite with $\xi_0 = \xi_{\min} = 0$. The proof of Theorem 5.1 requires delicate asymptotic analysis and is relegated to the Appendix. The FTP al-

gorithm based on SAVE or the directional regression can be developed parallel to SIR. Their screening consistency properties are left for future investigation.

To choose one model from the entire solution path $\mathcal{S} = \{\mathcal{S}^{(k)} : 1 \leq k \leq n\}$, we follow Chen and Chen (2008) and define the modified BIC criterion

$$\text{BIC}(\mathcal{F}) = -\log \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) \right\} + n^{-1}|\mathcal{F}|(\log n + 2 \log p). \quad (5.1)$$

The candidate model $\mathcal{S}^{(\hat{m})}$ is selected with $\hat{m} = \arg\min_{1 \leq k \leq n} \text{BIC}(\mathcal{S}^{(k)})$. The next result states that the selected model enjoys the screening consistency property.

Theorem 5.2. *Assume conditions (C1)-(C3) hold true. Then as $n \rightarrow \infty$ and $p \rightarrow \infty$, $Pr(\mathcal{A} \subset \mathcal{S}^{(\hat{m})}) \rightarrow 1$.*

Theorem 5.2 suggests we use BIC to determine the model size of the FTP algorithm in a data driven manner. The hybrid trace pursuit (HTP) algorithm combines FTP as the initial screening step and STP as the refined selection step. More specifically, the SIR-based HTP algorithm works as follows.

- (a) Perform SIR-based FTP and get solution path $\mathcal{S} = \{\mathcal{S}^{(k)} : 1 \leq k \leq n\}$.
- (b) Based on BIC criterion (5.1), select $\mathcal{S}^{(\hat{m})}$ with $\hat{m} = \arg\min_{1 \leq k \leq n} \text{BIC}(\mathcal{S}^{(k)})$.
- (c) Perform the SIR-based STP, where the full index set $\mathcal{I} = \{1, \dots, p\}$ is updated to the screened index set $\mathcal{S}^{(\hat{m})}$.

The HTP algorithms for SAVE and the directional regression can be implemented similarly, and the details are omitted.

6. Numerical studies

The proposed HTP algorithms are compared with existing model-free variable selection methods in this section. The screening performances of the FTP algorithms are evaluated as well.

6.1. Simulation studies

We consider the following models:

$$\begin{aligned} \text{I : } Y &= \text{sgn}(x_1 + x_p) \exp(x_2 + x_{p-1}) + \epsilon, \\ \text{II : } Y &= 2x_1^2 x_p^2 - 2x_2^2 x_{p-1}^2 + \epsilon, \\ \text{III : } Y &= x_1^4 - x_p^4 + 3 \exp(.8x_2 + .6x_{p-1}) + \epsilon. \end{aligned}$$

Unless specified otherwise, we set $\mathbf{X} = (x_1, \dots, x_p)^T$ to be multivariate normal with $E(\mathbf{X}) = \mathbf{0}$ and $\text{Var}(\mathbf{X}) = \Sigma$, and $\epsilon \sim N(0, \sigma^2)$ is independent of \mathbf{X} . The structural dimensions for Models I to III are respectively $q = 2, 4$ and 3 . The index set of significant predictors for all the three models is $\mathcal{A} = \{1, 2, p-1, p\}$. In all the simulation studies, we set $\sigma = .2$, the sample size $n = 300$, the number of slices $H = 4$. Consider three settings of p : $p = 10$ for small dimensionality, $p = 100$ for moderate dimensionality, and $p = 1000$ for high dimensionality. Denote the (i, j) th entry of Σ as $\rho^{|i-j|}$, and in the simulations, $\rho = 0$ is with uncorrelated predictors and $\rho = .5$ with correlated predictors.

When the SIR-based STP algorithm described in Section 4 is implemented, the threshold values \bar{c}^{SIR} and $\underline{c}^{\text{SIR}}$ cannot be easily determined as they depend on unknown rate ξ_{\min} relative to the sample size. Denote $D_{j|\mathcal{F}}^{\text{SIR}}$ as the weighted χ^2 distribution under $H_0 : Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$ in Theorem 3.1. It is easier in practice to choose quantiles of $D_{j|\mathcal{F}}^{\text{SIR}}$ as

the threshold values for the test statistics $T_{a_{\mathcal{F}}|\mathcal{F}}^{\text{SIR}}$ and $T_{d_{\mathcal{F}}|\{\mathcal{F}\setminus d_{\mathcal{F}}\}}^{\text{SIR}}$. Recall the definitions of $a_{\mathcal{F}}$ and $d_{\mathcal{F}}$ in (4.1) and (4.2). For the forward addition step, \mathcal{F} is updated to be $\mathcal{F} \cup a_{\mathcal{F}}$ if $T_{a_{\mathcal{F}}|\mathcal{F}}^{\text{SIR}} > D_{\alpha, a_{\mathcal{F}}|\mathcal{F}}^{\text{SIR}}$, the α th upper quantile of $D_{a_{\mathcal{F}}|\mathcal{F}}^{\text{SIR}}$. Similarly, in the backward deletion step, \mathcal{F} is updated to be $\mathcal{F} \setminus d_{\mathcal{F}}$ if $T_{d_{\mathcal{F}}|\{\mathcal{F}\setminus d_{\mathcal{F}}\}}^{\text{SIR}} < D_{\alpha, d_{\mathcal{F}}|\{\mathcal{F}\setminus d_{\mathcal{F}}\}}^{\text{SIR}}$, the α th upper quantile of $D_{d_{\mathcal{F}}|\{\mathcal{F}\setminus d_{\mathcal{F}}\}}^{\text{SIR}}$. Jiang and Liu (2013) suggest trying α over the grid points in the interval $(0, 1)$, and determining the final α by the cross validation. For ease of implementation, we set the level of α as $.1p^{-1}$ in all the simulation studies. We follow Bentler and Xie (2000) to approximate the α th upper quantile of a weighted χ^2 distribution. Other approximations, such as Field (1993), Cook and Setodji (2003), can be used as well. The STP algorithms based on SAVE and directional regression are carried out in a similar fashion.

We examine the performances of the HTP algorithms for variable selection in Tables 1 to 5. The HTP algorithms that are based on SIR, SAVE and the directional regression are denoted by HTP-SIR, HTP-SAVE and HTP-DR respectively. Based on the $N = 100$ repetitions, we report the underfitted count (UF), the correctly fitted count (CF), the overfitted count (OF), and the average model size (MS). Let $\hat{\mathcal{A}}_{(i)}$ be the estimated active set in the i th repetition and define

$$\begin{aligned} UF &= \sum_{i=1}^N I(\mathcal{A} \not\subseteq \hat{\mathcal{A}}_{(i)}), CF = \sum_{i=1}^N I(\mathcal{A} = \hat{\mathcal{A}}_{(i)}), \\ OF &= \sum_{i=1}^N I(\mathcal{A} \subset \hat{\mathcal{A}}_{(i)}), \text{ and } MS = N^{-1} \sum_{i=1}^N |\hat{\mathcal{A}}_{(i)}|. \end{aligned}$$

The selection performance of Model I is summarized in Table 1. This model favors SIR as Y is monotone of the two linear combinations $x_1 + x_p$ and $x_2 + x_{p-1}$. HTP-SIR works very well for this model, as condition (4.3) is satisfied here. The performance of

Table 1: Comparison among three HTP algorithms for Model I.

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
I	HTP-SIR	10	0	100	0	4.00	0	100	0	4.00
		100	0	100	0	4.00	0	100	0	4.00
		1000	0	100	0	4.00	0	100	0	4.00
	HTP-SAVE	10	9	59	32	4.31	4	39	57	4.00
		100	32	0	68	20.53	46	1	53	18.14
		1000	90	0	10	18.93	91	0	9	15.59
	HTP-DR	10	0	98	2	4.02	0	99	1	4.01
		100	0	95	5	4.07	0	93	7	4.08
		1000	0	96	4	4.04	0	94	6	4.08

Table 2: Comparison among three HTP algorithms for Model II.

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
II	HTP-SIR	10	100	0	0	.31	100	0	0	.24
		100	100	0	0	.13	100	0	0	.08
		1000	100	0	0	.03	100	0	0	.03
	HTP-SAVE	10	2	97	1	3.99	2	94	4	4.02
		100	3	53	44	4.63	3	50	47	4.71
		1000	3	48	49	4.79	7	41	52	4.95
	HTP-DR	10	3	95	2	3.99	3	93	4	4.01
		100	2	56	42	4.70	3	46	51	4.77
		1000	7	44	49	4.76	7	45	48	4.91

HTP-SIR keeps up with diverging p , which validates our theoretical finding in Theorem 5.2. We know from the sufficient dimension reduction literature that SAVE is not efficient when predictors are linked to the response through monotone functions. We see that HTP-SAVE has very unstable performances, which either underfits or overfits with a large probability. HTP-DR performs similarly to HTP-SIR, and fits correctly with a dominant probability.

Table 2 reports the performance of HTP methods for Model II, which favors SAVE as Y depends on quadratic functions x_1^2, x_2^2, x_{p-1}^2 and x_p^2 . HTP-SAVE works reasonably well for this model. It correctly recovers \mathcal{A} with a dominant probability when $p = 10$.

Table 3: Comparison among three HTP algorithms for Model III.

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
III	HTP-SIR	10	100	0	0	2.07	100	0	0	2.23
		100	100	0	0	2.58	100	0	0	3.56
		1000	100	0	0	6.34	100	0	0	6.56
	HTP-SAVE	10	4	33	63	5.14	7	45	48	4.61
		100	47	8	45	12.42	36	6	58	8.43
		1000	86	0	14	21.76	78	2	20	16.86
	HTP-DR	10	0	91	9	4.11	0	98	2	4.02
		100	3	83	14	4.13	4	79	17	4.16
		1000	4	88	8	4.06	5	61	34	4.39

With probability close to one, HTP-SAVE either correctly identifies or overfits \mathcal{A} when $p = 100$ or $p = 1000$. The average model size of HTP-SAVE for Model II is not much larger than 4, indicating very mild overfitting. From the average model size, we see that HTP-SIR underfits and misses all four variables with high probability. This is as expected because condition (4.3) is violated for this model. HTP-DR has very similar performance to HTP-SAVE for this model.

The comparison of Model III is reported in Table 3. Model III involves quartic functions x_1^4, x_p^4 , as well as a monotone function $3 \exp(.8x_2 + .6x_{p-1})$, and is thus favorable for the directional regression. Both HTP-SIR and HTP-SAVE would fail. HTP-SIR misses two variables on average, which should be x_1 and x_p involved in the two quartic terms. As we have seen in Table 1, HTP-SAVE either underfits or overfits with a large probability due to the monotone function $3 \exp(.8x_2 + .6x_{p-1})$. HTP-DR still enjoys good performance for Model III, and correctly recovers \mathcal{A} with a large probability. The average model size is always close to 4, indicating a good overall fit.

To check the performances of the HTP algorithms for nonnormal predictors, consider $\mathbf{X} = (x_1, \dots, x_p)^T$ with : case (i), $x_i \sim \text{Uniform}(1, 2)$; case (ii), $x_i \sim \text{Exponential}(1)$; case

Table 4: Comparison among three HTP algorithms for Model III with nonnormal \mathbf{X} .

$p = 1000$	case (i), Uniform				case (ii), Exponential				case (iii), Geometric			
Method	UF	CF	OF	MS	UF	CF	OF	MS	UF	CF	OF	MS
HTP-SIR	0	88	12	4.12	0	99	1	4.01	1	98	1	4.00
HTP-SAVE	0	25	75	5.62	7	72	19	4.39	27	17	56	5.58
HTP-DR	0	93	7	4.08	5	81	14	4.06	11	86	3	3.82

(iii), $x_i \sim \text{Geometric}(.5)$. In all three cases, x_i is independent of x_j for $i \neq j$, $1 \leq i, j \leq p$. We focus on Model III with $n = 300$ and $p = 1000$, and report the results in Table 4. We see that HTP-DR with nonnormal predictors has similar performance compared to its counterpart with normal predictors in Table 3. HTP-SAVE has unstable performance as before. The performance of HTP-SIR with nonnormal predictors actually has significant improvement over its counterpart with normal predictors in Table 3. We have seen before that SIR-based method can not pick up quartic terms x_1^4 , x_p^4 involved in Model III with $x_i \sim N(0, 1)$. This happens because the symmetry of the x_i distribution coincides with the symmetry of the link function x_i^4 . Since the distribution of the nonnormal x_i is no longer symmetric about 0, HTP-SIR is able to select x_1 and x_p in the quartic terms. We conclude from Table 4 that the proposed HTP algorithms are not sensitive to the normality assumption of the predictors.

Next we focus on the challenging case of $p = 1000$, and compare the performances of HTP-DR with existing methods in Table 5. Only COP (Zhong et al., 2012) and SIRI (Jiang and Liu, 2013) are included in the comparison, as other methods such as CISE (Chen et al., 2010) can not handle $p > n$. The R codes for COP and SIRI are made available by the respective authors. COP works well for Model I, and underfits for Models II and III. COP has similar performances to HTP-SIR as both the methods are based on SIR. SIRI works well for Models I and III, and is likely to underfit for Model II. We suspect the relatively large structure dimension $q = 4$ in Model II is a probable

Table 5: Comparison between COP, SIRI and HTP-DR. Selection performances based on $p = 1000$ and $N = 100$ repetitions are reported.

Model	Method	$\rho = 0$				$\rho = .5$			
		UF	CF	OF	MS	UF	CF	OF	MS
I	COP	0	86	14	4.14	0	85	15	4.16
	SIRI	0	66	34	4.46	0	86	14	4.19
	HTP-DR	0	96	4	4.04	0	94	6	4.08
II	COP	100	0	0	4.00	100	0	0	4.00
	SIRI	52	38	10	3.79	36	45	19	4.05
	HTP-DR	7	44	49	4.76	7	45	48	4.91
III	COP	100	0	0	3.09	100	0	0	3.15
	SIRI	1	99	0	3.99	2	98	0	3.98
	HTP-DR	4	88	8	4.06	5	61	34	4.39

Table 6: Comparison between SIS, DC-SIS and FTP algorithms for screening. Frequencies of cases including all active predictors are reported based on $p = 2000$ and $N = 100$ repetitions.

Method	$\rho = 0$			$\rho = .5$		
	Model I	Model II	Model III	Model I	Model II	Model III
SIS	15	0	3	97	1	23
DC-SIS	100	100	100	100	100	100
FTP-SIR	100	0	0	100	0	0
FTP-SAVE	12	97	7	10	98	31
FTP-DR	100	97	98	100	98	97

cause for the deficiency of SIRI. HTP-DR completely avoids estimating the structure dimension q , and has decent overall performance.

To compare the screening performances of the FTP algorithms, we report in Table 6 the frequencies in $N = 100$ repetitions when all the significant predictors are included after screening. The FTP algorithms that are based on SIR, SAVE and the directional regression are denoted by FTP-SIR, FTP-SAVE and FTP-DR respectively. The sure independence screening (SIS) in Fan and Lv (2008) and the distance correlation-based SIS (DC-SIS) (Li et al., 2012) are also included. The model size of SIS and DC-SIS

Table 7: Classification results based on LDA for the leukemia data.

Method	COP	SIRI	HTP-SIR	HTP-SAVE	HTP-DR
Training error counts	0	1	2	NA	2
Testing error counts	2	1	2	NA	1
Number of genes selected	2	6	1	0	2

is set to be $\lceil n/\log n \rceil$, while the model size of FTP is determined by the BIC criterion in (5.1). SIS does not work well as it is designed for linear models. FTP-SIR works well for Model I as the significant predictors appear in the monotone link functions. FTP-SAVE works well for Model II as the significant predictors appear in the quadratic link functions. FTP-DR performs similarly to the state of the art method DC-SIS, and retains all active predictors with probability close to one across all the three models. In addition, the BIC in (5.1) for FTP-DR leads to average model size of 20, which is much smaller compared to $\lceil 300/\log 300 \rceil = 53$, the model size of DC-SIS.

6.2. Real data analysis

We consider the leukemia data from the high-density Affymetrix oligonucleotide arrays (Golub et al, 1999). This data set has become a benchmark in many gene expression studies. See, for example, Dettling (2004). There are 38 training samples and 34 testing samples, with 3571 genes in each sample. The response is 0 or 1 describing two subtypes of leukemia. We first perform variable selection that is based on the training set, build a classification rule with the linear discriminant analysis (LDA), and then apply this rule to the testing set. We compare the classification results together with the number of genes selected in Table 7. HTP-SAVE fails to select any significant gene, suggesting that the subtypes of leukemia may depend on the genes through some monotone link functions. COP, SIRI, HTP-SIR and HTP-DR all lead to similar classification performances. While both SIRI and HTP-DR have the smallest testing error count, HTP-DR

needs only 2 genes compared to 6 genes selected by SIRI.

7. Discussions

For high dimensional data with unknown link functions between predictors and response, it is desirable to perform variable selection in a model-free fashion. We have proposed a versatile framework for variable selection via stepwise trace pursuit, which can be viewed as a model-free counterpart of the classical stepwise regression. An important connection between sufficient dimension reduction and model-free variable selection is revealed in Cook (2004) via the marginal coordinate test. However, it is not applicable when p is larger than n . Stepwise trace pursuit provides the missing link between sufficient dimension reduction and model-free variable selection in the high dimensional settings. While our discussions in this paper are based on SIR, SAVE and the directional regression, the general principle of trace pursuit allows its extension to other sufficient dimension reduction methods as well.

As an important preprocessing step for ultrahigh dimensional data, variable screening is first proposed in Fan and Lv (2008) and has received much attention in the recent literature (Zhu et al., 2011; Li et al., 2012; He et al., 2013; Chang et al., 2013, Lin et al 2013). Forward trace pursuit is introduced in this paper for model-free variable screening under the sufficient dimension reduction framework. The screening consistency property of forward regression in linear models is established in Wang (2009), which is extended to model-free setting via SIR-based forward trace pursuit. The theoretical properties of forward trace pursuit approaches that are based on other sufficient dimension reduction methods warrant future investigation.

8. Appendix: Proofs

8.1. Proofs of Theorems in Section 2

PROOF OF PROPOSITION 2.1. Let $\lambda_1 \geq \dots \geq \lambda_q$ be the q nonzero eigenvalues of \mathbf{M}^{SIR} . Denote $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_q$ as the corresponding eigenvectors. Let $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}_i$ for $i = 1, \dots, q$. Note that $\mathbf{M}^{\text{SIR}} = \sum_{i=1}^q \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T = \boldsymbol{\Sigma}^{1/2} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \right) \boldsymbol{\Sigma}^{1/2}$. Thus we have

$$\text{tr}(\mathbf{M}^{\text{SIR}}) = \text{tr} \left\{ \boldsymbol{\Sigma} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \right) \right\}. \quad (8.1)$$

The LCM assumption (2.1) guarantees that $\boldsymbol{\beta}_i \in \mathcal{S}_{Y|\mathbf{X}}$. For $\boldsymbol{\beta}_i = (\beta_{i,1}, \dots, \beta_{i,p})^T$, let $\boldsymbol{\beta}_{i,\mathcal{A}} = \{\beta_{i,j} : j \in \mathcal{A}\}$ and $\boldsymbol{\beta}_{i,\mathcal{A}^c} = \{\beta_{i,j} : j \in \mathcal{A}^c\}$. The fact that $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$ and the definition of $\mathcal{S}_{Y|\mathbf{X}}$ together imply that $\boldsymbol{\beta}_{i,\mathcal{A}^c} = \mathbf{0}$. Hence (8.1) becomes

$$\text{tr}(\mathbf{M}^{\text{SIR}}) = \text{tr} \left\{ \boldsymbol{\Sigma}_{\mathcal{A}} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_{i,\mathcal{A}} \boldsymbol{\beta}_{i,\mathcal{A}}^T \right) \right\}, \quad (8.2)$$

where $\mathbf{X}_{\mathcal{A}} = \{x_i : i \in \mathcal{A}\}$ and $\boldsymbol{\Sigma}_{\mathcal{A}} = \text{Var}(\mathbf{X}_{\mathcal{A}})$. By definition, we have

$$\text{tr}(\mathbf{M}_{\mathcal{A}}^{\text{SIR}}) = \text{tr}\{\text{Cov}(E(\mathbf{Z}_{\mathcal{A}}|Y))\} = \text{tr}\{\boldsymbol{\Sigma}_{\mathcal{A}}^{-1} \text{Cov}(E(\mathbf{X}_{\mathcal{A}}|Y))\}. \quad (8.3)$$

Assume $\mathcal{A} = \{1, 2, \dots, K\}$ without loss of generality. Note that

$$\begin{aligned} \text{Cov}(E(\mathbf{X}|Y)) &= \boldsymbol{\Sigma}^{1/2} \mathbf{M}^{\text{SIR}} \boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \right) \boldsymbol{\Sigma} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A}^c} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^q \lambda_i \boldsymbol{\beta}_{i,\mathcal{A}} \boldsymbol{\beta}_{i,\mathcal{A}}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A}^c} \end{pmatrix}, \end{aligned}$$

the upper left block of which implies that $\text{Cov}(E(\mathbf{X}_{\mathcal{A}}|Y)) = \boldsymbol{\Sigma}_{\mathcal{A}} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_{i,\mathcal{A}} \boldsymbol{\beta}_{i,\mathcal{A}}^T \right) \boldsymbol{\Sigma}_{\mathcal{A}}$.

Plug it into (8.3), and we get $\text{tr}(\mathbf{M}_{\mathcal{A}}^{\text{SIR}}) = \text{tr}\left\{\Sigma_{\mathcal{A}}\left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_{i,\mathcal{A}} \boldsymbol{\beta}_{i,\mathcal{A}}^T\right)\right\}$. Together with (8.2), we have $\text{tr}(\mathbf{M}_{\mathcal{A}}^{\text{SIR}}) = \text{tr}(\mathbf{M}^{\text{SIR}})$. For \mathcal{F} satisfying $\mathcal{A} \subseteq \mathcal{F}$, the proof is similar and omitted. \square

PROOF OF THEOREM 2.1. Condition (2.5) implies that for $\mathcal{F} \subset \mathcal{I}$ and $j \in \mathcal{F}^c$, we have $E(x_j|\mathbf{X}_{\mathcal{F}}) = E^T(x_j \mathbf{X}_{\mathcal{F}}) \Sigma_{\mathcal{F}}^{-1} \mathbf{X}_{\mathcal{F}}$. Recall that $|\mathcal{F}|$ denotes the cardinality of \mathcal{F} . For the first part, define $(|\mathcal{F}| + 1) \times (|\mathcal{F}| + 1)$ dimensional matrices \mathbf{A} and \mathbf{C} as

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \Sigma_{\mathcal{F}}^{-1} & 1 \end{pmatrix} \text{ and } \mathbf{C} = \begin{pmatrix} \Sigma_{\mathcal{F}} & \mathbf{0} \\ \mathbf{0} & \sigma_{j|\mathcal{F}}^2 \end{pmatrix}. \quad (8.4)$$

Recall that $\mathbf{U}_h = E(\mathbf{X}|Y \in J_h)$. It is easy to check that

$$\mathbf{A} \mathbf{X}_{\mathcal{F} \cup j} = \begin{pmatrix} \mathbf{X}_{\mathcal{F}} \\ x_{j|\mathcal{F}} \end{pmatrix} \text{ and } \mathbf{A} \mathbf{U}_{\mathcal{F} \cup j, h} = \begin{pmatrix} \mathbf{U}_{\mathcal{F}, h} \\ E(x_{j|\mathcal{F}}|Y \in J_h) \end{pmatrix}. \quad (8.5)$$

Because $E(\mathbf{X}_{\mathcal{F}} x_{j|\mathcal{F}}) = E\{\mathbf{X}_{\mathcal{F}} E(x_{j|\mathcal{F}}|\mathbf{X}_{\mathcal{F}})\} = 0$, we have $\text{Var}(\mathbf{A} \mathbf{X}_{\mathcal{F} \cup j}) = \mathbf{A} \Sigma_{\mathcal{F} \cup j} \mathbf{A}^T = \mathbf{C}$. Hence $\Sigma_{\mathcal{F} \cup j}^{-1} = \mathbf{A}^T \mathbf{C}^{-1} \mathbf{A}$. Together with $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) = \text{tr}\left\{\Sigma_{\mathcal{F} \cup j}^{-1} \left(\sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F} \cup j, h} \mathbf{U}_{\mathcal{F} \cup j, h}^T\right)\right\}$, we get

$$\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) = \text{tr}\left(\mathbf{C}^{-1} \left\{\sum_{h=1}^H p_h (\mathbf{A} \mathbf{U}_{\mathcal{F} \cup j, h}) (\mathbf{A} \mathbf{U}_{\mathcal{F} \cup j, h})^T\right\}\right). \quad (8.6)$$

Plug (8.4) and (8.5) into (8.6), and we get

$$\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) = \text{tr}\left\{\Sigma_{\mathcal{F}}^{-1} \left(\sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F}, h} \mathbf{U}_{\mathcal{F}, h}^T\right)\right\} + \sum_{h=1}^H p_h \sigma_{j|\mathcal{F}}^{-2} E^2(x_{j|\mathcal{F}}|Y \in J_h).$$

It follows immediately that $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = \sum_{h=1}^H p_h \gamma_{j|\mathcal{F}, h}^2$.

For the second part, note that $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$, $\mathcal{A} \subseteq \mathcal{F}$ and $j \in \mathcal{F}^c$ together imply that $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$. Thus we have $E(x_j | Y, \mathbf{X}_{\mathcal{F}}) = E(x_j | \mathbf{X}_{\mathcal{F}})$. It follows that

$$E(x_j | Y) = E\{E(x_j | Y, \mathbf{X}_{\mathcal{F}}) | Y\} = E\{E(x_j | \mathbf{X}_{\mathcal{F}}) | Y\}. \quad (8.7)$$

As a result $E(x_j | \mathcal{F} | Y) = E[\{x_j - E(x_j | \mathbf{X}_{\mathcal{F}})\} | Y] = 0$ and $\gamma_{j|\mathcal{F},h} = 0$. Hence $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = 0$. \square

PROOF OF THEOREM 2.2. Define \mathbf{A} and \mathbf{C} as in (8.4). Denote $\mathbf{B}_{\mathcal{F},h} = \Sigma_{\mathcal{F}} - \mathbf{V}_{\mathcal{F},h} + \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T$. It follows that

$$\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SAVE}}) = \sum_{h=1}^H p_h \text{tr}(\Sigma_{\mathcal{F} \cup j}^{-1} \mathbf{B}_{\mathcal{F} \cup j,h} \Sigma_{\mathcal{F} \cup j}^{-1} \mathbf{B}_{\mathcal{F} \cup j,h}).$$

By noticing $\text{Var}(\mathbf{X}_{\mathcal{F} \cup j}) = \Sigma_{\mathcal{F} \cup j}$ and $\text{Var}(\mathbf{A} \mathbf{X}_{\mathcal{F} \cup j}) = \mathbf{C}$, we have $\Sigma_{\mathcal{F} \cup j}^{-1} = \mathbf{A}^T \mathbf{C}^{-1} \mathbf{A}$. Let $\mathbf{D}_h = \mathbf{C}^{-1/2} \mathbf{A} \mathbf{B}_{\mathcal{F} \cup j,h} \mathbf{A}^T \mathbf{C}^{-1/2}$. Then

$$\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SAVE}}) = \sum_{h=1}^H p_h \text{tr}(\mathbf{C}^{-1} \mathbf{A} \mathbf{B}_{\mathcal{F} \cup j,h} \mathbf{A}^T \mathbf{C}^{-1} \mathbf{A} \mathbf{B}_{\mathcal{F} \cup j,h} \mathbf{A}^T) = \sum_{h=1}^H p_h \text{tr}(\mathbf{D}_h \mathbf{D}_h).$$

To calculate $\mathbf{A} \mathbf{B}_{\mathcal{F} \cup j,h} \mathbf{A}^T = \mathbf{A} (\Sigma_{\mathcal{F} \cup j} - \mathbf{V}_{\mathcal{F} \cup j,h} + \mathbf{U}_{\mathcal{F} \cup j,h} \mathbf{U}_{\mathcal{F} \cup j,h}^T) \mathbf{A}^T$, note that

$$\begin{aligned} \mathbf{A} \Sigma_{\mathcal{F} \cup j} \mathbf{A}^T &= \begin{pmatrix} \Sigma_{\mathcal{F}} & \mathbf{0} \\ \mathbf{0} & \sigma_{j|\mathcal{F}}^2 \end{pmatrix}, \mathbf{A} \mathbf{V}_{\mathcal{F} \cup j,h} \mathbf{A}^T = \begin{pmatrix} \mathbf{V}_{\mathcal{F},h} & E(\mathbf{X}_{\mathcal{F}} x_j | Y \in J_h) \\ E^T(\mathbf{X}_{\mathcal{F}} x_j | Y \in J_h) & E(x_j^2 | Y \in J_h) \end{pmatrix}, \\ \text{and } \mathbf{A} \mathbf{U}_{\mathcal{F} \cup j,h} \mathbf{U}_{\mathcal{F} \cup j,h}^T \mathbf{A}^T &= \begin{pmatrix} \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T & \mathbf{U}_{\mathcal{F},h} E(x_j | Y \in J_h) \\ \mathbf{U}_{\mathcal{F},h}^T E(x_j | Y \in J_h) & E^2(x_j | Y \in J_h) \end{pmatrix}. \end{aligned}$$

Let $\psi_{j|\mathcal{F},h} = \mathbf{U}_{\mathcal{F},h} E(x_j | Y \in J_h) - E(\mathbf{X}_{\mathcal{F}} x_j | Y \in J_h)$ and $D_{22} = \sigma_{j|\mathcal{F}}^2 - E(x_j^2 | Y \in J_h)$.

$J_h) + E^2(x_{j|\mathcal{F}}|Y \in J_h)$. Then

$$\mathbf{A}\mathbf{B}_{\mathcal{F} \cup j, h}\mathbf{A}^T = \begin{pmatrix} \mathbf{B}_{\mathcal{F}, h} & \boldsymbol{\psi}_{j|\mathcal{F}, h} \\ \boldsymbol{\psi}_{j|\mathcal{F}, h}^T & D_{22} \end{pmatrix}$$

It follows that $\mathbf{D}_h = \mathbf{C}^{-1/2}\mathbf{A}\mathbf{B}_{\mathcal{F} \cup j, h}\mathbf{A}^T\mathbf{C}^{-1/2}$ becomes

$$\mathbf{D}_h = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}\mathbf{B}_{\mathcal{F}, h}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} & \boldsymbol{\phi}_{j|\mathcal{F}, h} \\ \boldsymbol{\phi}_{j|\mathcal{F}, h}^T & 1 - \zeta_{j|\mathcal{F}, h} + \gamma_{j|\mathcal{F}, h}^2 \end{pmatrix}.$$

The conclusion of the first part is then obvious.

For the second part, we now show that $\boldsymbol{\phi}_{j|\mathcal{F}, h} = \mathbf{0}$, $\gamma_{j|\mathcal{F}, h} = 0$, and $\zeta_{j|\mathcal{F}, h} = 1$. Note that $\gamma_{j|\mathcal{F}, h} = 0$ in the proof of Theorem 2.1. Similar to (8.7) where we have shown $E(x_j|Y) = E\{E(x_j|\mathbf{X}_{\mathcal{F}})|Y\}$, it can be shown that $E(x_j\mathbf{X}_{\mathcal{F}}|Y) = E\{\mathbf{X}_{\mathcal{F}}E(x_j|\mathbf{X}_{\mathcal{F}})|Y\}$. Hence $E(x_j|Y) = 0$ and $E(\mathbf{X}_{\mathcal{F}}x_{j|\mathcal{F}}|Y) = \mathbf{0}$. It follows that $\text{Cov}(\mathbf{X}_{\mathcal{F}}, x_{j|\mathcal{F}}|Y) = \mathbf{0}$, and $\boldsymbol{\phi}_{j|\mathcal{F}, h} = -\boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}\text{Cov}(\mathbf{X}_{\mathcal{F}}, x_{j|\mathcal{F}}|Y \in J_h) = \mathbf{0}$. It remains to show that $\zeta_{j|\mathcal{F}, h} = E(\gamma_{j|\mathcal{F}}^2|Y \in J_h) = 1$, or $E(x_{j|\mathcal{F}}^2|Y) = \text{Var}(x_{j|\mathcal{F}})$. Because $x_{j|\mathcal{F}} = x_j - E(x_j|\mathbf{X}_{\mathcal{F}})$ and $\text{Cov}\{x_j, E(x_j|\mathbf{X}_{\mathcal{F}})\} = E\{x_j E(x_j|\mathbf{X}_{\mathcal{F}})\} = \text{Var}\{E(x_j|\mathbf{X}_{\mathcal{F}})\}$, we have

$$\text{Var}(x_{j|\mathcal{F}}) = \text{Var}(x_j) - \text{Var}\{E(x_j|\mathbf{X}_{\mathcal{F}})\} = E\{\text{Var}(x_j|\mathbf{X}_{\mathcal{F}})\}. \quad (8.8)$$

The subset CCV condition (2.6) implies that $E\{(x_j - E(x_j|\mathbf{X}_{\mathcal{F}}))^2|Y, \mathbf{X}_{\mathcal{F}}\} = E\{(x_j - E(x_j|\mathbf{X}_{\mathcal{F}}))^2|\mathbf{X}_{\mathcal{F}}\} = \text{Var}(x_j|\mathbf{X}_{\mathcal{F}})$. Thus we have

$$E(x_{j|\mathcal{F}}^2|Y) = E\{(x_j - E(x_j|\mathbf{X}_{\mathcal{F}}))^2|Y, \mathbf{X}_{\mathcal{F}}\} = E\{\text{Var}(x_j|\mathbf{X}_{\mathcal{F}})|Y\}. \quad (8.9)$$

Compare (8.8) with (8.9) and we get the desired result. \square

PROOF OF THEOREM 2.3. Let $\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR1}} = \sum_{h=1}^H p_h \left(\mathbf{I}_{|\mathcal{F}|+1} - \boldsymbol{\Sigma}_{\mathcal{F} \cup j}^{-1/2} \mathbf{V}_{\mathcal{F} \cup j, h} \boldsymbol{\Sigma}_{\mathcal{F} \cup j}^{-1/2} \right)^2$, $\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR2}} = \sum_{h=1}^H p_h \boldsymbol{\Sigma}_{\mathcal{F} \cup j}^{-1/2} \mathbf{U}_{\mathcal{F} \cup j, h} \mathbf{U}_{\mathcal{F} \cup j, h}^T \boldsymbol{\Sigma}_{\mathcal{F} \cup j}^{-1/2}$, and $m_{\mathcal{F} \cup j}^{\text{DR3}} = \sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F} \cup j, h}^T \boldsymbol{\Sigma}_{\mathcal{F} \cup j}^{-1} \mathbf{U}_{\mathcal{F} \cup j, h}$. Then $\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR}}/2$ can be written as

$$\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR1}} + (\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR2}})^2 + m_{\mathcal{F} \cup j}^{\text{DR3}} \mathbf{M}_{\mathcal{F} \cup j}^{\text{DR2}}. \quad (8.10)$$

The first term in (8.10) can be shown to satisfy

$$\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR1}}) = \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{DR1}}) + \sum_{h=1}^H p_h \left\{ (1 - \zeta_{j|\mathcal{F}, h})^2 + 2\boldsymbol{\nu}_{j|\mathcal{F}, h}^T \boldsymbol{\nu}_{j|\mathcal{F}, h} \right\}.$$

The second term in (8.10) can be shown to satisfy

$$\text{tr} \left\{ (\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR2}})^2 \right\} = \text{tr} \left\{ (\mathbf{M}_{\mathcal{F}}^{\text{DR2}})^2 \right\} + 2\boldsymbol{\iota}_{j|\mathcal{F}}^T \boldsymbol{\iota}_{j|\mathcal{F}} + \varrho_{j|\mathcal{F}}^2.$$

The last term in (8.10) can be shown to satisfy

$$\text{tr} (m_{\mathcal{F} \cup j}^{\text{DR3}} \mathbf{M}_{\mathcal{F} \cup j}^{\text{DR2}}) = \text{tr} (m_{\mathcal{F}}^{\text{DR3}} \mathbf{M}_{\mathcal{F}}^{\text{DR2}}) + 2\kappa_{\mathcal{F}} \varrho_{j|\mathcal{F}} + \varrho_{j|\mathcal{F}}^2.$$

Together we get the first part of Theorem 2.3. For the second part, we have seen in the proof of Theorem 2.2 that $\gamma_{j|\mathcal{F}, h} = 0$, and $\zeta_{j|\mathcal{F}, h} = 1$ given that $\mathcal{A} \subseteq \mathcal{F}$. It is easy to see that $\boldsymbol{\nu}_{j|\mathcal{F}, h} = \mathbf{0}$, $\boldsymbol{\iota}_{j|\mathcal{F}} = \mathbf{0}$ and $\varrho_{j|\mathcal{F}} = 0$. The conclusion is then obvious. \square

8.2. Proofs of Theorems in Section 3

We use Frechet derivative representation to derive the asymptotic distributions of

$T_{j|\mathcal{F}}^{\text{SIR}}$, $T_{j|\mathcal{F}}^{\text{SAVE}}$ and $T_{j|\mathcal{F}}^{\text{DR}}$. Let F be the joint distribution of (\mathbf{X}, Y) and F_n be the empirical distribution based on the i.i.d. sample $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. Let G be a real or matrix valued functional. Then $G(F_n)$ has the following asymptotic expansion under regularity conditions,

$$G(F_n) = G(F) + E_n\{G^*(F)\} + O_p(n^{-1}), \quad (8.11)$$

where $G(F)$ is nonrandom, and $E_n\{G^*(F)\} = O_p(n^{-1/2})$ as $G^*(F)$ satisfies $E\{G^*(F)\} = 0$. Please refer to Fernholz (1983) for more details about Frechet derivative and the regularity conditions.

Recall the following definitions involved in Theorem 3.1: $R_h = I(Y \in J_h)$, $p_h = E(I(Y \in J_h))$, $\Sigma_{\mathcal{F}} = \text{Var}(\mathbf{X}_{\mathcal{F}})$, $\mathbf{U}_{\mathcal{F},h} = E(\mathbf{X}_{\mathcal{F}}|Y \in J_h)$, $x_{j|\mathcal{F}} = x_j - E(x_j|\mathbf{X}_{\mathcal{F}})$, $\sigma_{j|\mathcal{F}}^2 = \text{Var}(x_{j|\mathcal{F}})$, $\gamma_{j|\mathcal{F}} = x_{j|\mathcal{F}}/\sigma_{j|\mathcal{F}}$, and $\gamma_{j|\mathcal{F},h} = E(\gamma_{j|\mathcal{F}}|Y \in J_h)$. Denote $u_{j,h} = E(x_j|Y \in J_h)$ and $\boldsymbol{\vartheta}_{j|\mathcal{F}} = \Sigma_{\mathcal{F}}^{-1}E(x_j\mathbf{X}_{\mathcal{F}})$. Then for $\mathcal{F} \subset \mathcal{I}$ and $j \in \mathcal{F}^c$, condition (2.5) implies $E(x_j|\mathbf{X}_{\mathcal{F}}) = \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{X}_{\mathcal{F}}$.

Lemma 1. *Assume the conditions of Theorem 3.1 are satisfied. Under $H_0 : Y \perp\!\!\!\perp x_j|\mathbf{X}_{\mathcal{F}}, j \in \mathcal{F}^c$, the expansions of $\hat{\Sigma}_{\mathcal{F}}$, $\hat{\Sigma}_{\mathcal{F}}^{-1}$, $\hat{\mathbf{U}}_{\mathcal{F},h}$, $\hat{\boldsymbol{\vartheta}}_{j|\mathcal{F}}$, and $\hat{\gamma}_{j|\mathcal{F},h}$ have the form (8.11), where we replace $G(F)$ with $\Sigma_{\mathcal{F}}$, $\Sigma_{\mathcal{F}}^{-1}$, $\mathbf{U}_{\mathcal{F},h}$, $\boldsymbol{\vartheta}_{j|\mathcal{F}}$, or $\gamma_{j|\mathcal{F},h}$, and we replace $G^*(F)$ with $\Sigma_{\mathcal{F}}^* = \mathbf{X}_{\mathcal{F}}\mathbf{X}_{\mathcal{F}}^T - \Sigma_{\mathcal{F}}$, $(\Sigma_{\mathcal{F}}^{-1})^* = -\Sigma_{\mathcal{F}}^{-1}\Sigma_{\mathcal{F}}^*\Sigma_{\mathcal{F}}^{-1}$, $\mathbf{U}_{\mathcal{F},h}^* = (\mathbf{X}_{\mathcal{F}} - \mathbf{U}_{\mathcal{F},h})R_h/p_h - \mathbf{X}_{\mathcal{F}}$, $\boldsymbol{\vartheta}_{j|\mathcal{F}}^* = \Sigma_{\mathcal{F}}^{-1}\{x_j\mathbf{X}_{\mathcal{F}} - E(x_j\mathbf{X}_{\mathcal{F}})\} + (\Sigma_{\mathcal{F}}^{-1})^*E(x_j\mathbf{X}_{\mathcal{F}})$, or $\gamma_{j|\mathcal{F},h}^* = \{u_{j,h}^* - (\boldsymbol{\vartheta}_{j|\mathcal{F}}^*)^T\mathbf{U}_{\mathcal{F},h} - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T\mathbf{U}_{\mathcal{F},h}^*\}/\sigma_{j|\mathcal{F}}$.*

PROOF OF LEMMA 1. The expansions of $\hat{\Sigma}_{\mathcal{F}}$, $\hat{\Sigma}_{\mathcal{F}}^{-1}$, $\hat{\mathbf{U}}_{\mathcal{F},h}$ and $\boldsymbol{\vartheta}_{j|\mathcal{F}}$ are similar to those in Lemma 1 in Li and Wang (2007), and thus omitted. With condition (2.5), we have

$x_{j|\mathcal{F}} = x_j - E(x_j|\mathbf{X}_{\mathcal{F}}) = x_j - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{X}_{\mathcal{F}}$. For the expansion of $\hat{\gamma}_{j|\mathcal{F},h}$, notice that

$$\sigma_{j|\mathcal{F}} \gamma_{j|\mathcal{F},h} = E(x_j - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{X}_{\mathcal{F}} | Y \in J_h) = u_{j,h} - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{U}_{\mathcal{F},h}.$$

We have seen in the proof of Theorem 2.1 that $\gamma_{j|\mathcal{F},h} = 0$ with $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$ and condition (2.5). Taking Frechet derivative on both sides of the listed equation above, we get

$$\sigma_{j|\mathcal{F}} \gamma_{j|\mathcal{F},h}^* = u_{j,h}^* - (\boldsymbol{\vartheta}_{j|\mathcal{F}}^*)^T \mathbf{U}_{\mathcal{F},h} - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{U}_{\mathcal{F},h}^*. \quad \square$$

Note that $u_{j,h}$ is a special case of $\mathbf{U}_{\mathcal{F},h}$ with \mathcal{F} replaced by j . Thus expansion of $\hat{u}_{j,h}$ follows the same form of $\hat{\mathbf{U}}_{\mathcal{F},h}$ and is omitted.

PROOF OF THEOREM 3.1. Let $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SIR}} = (\hat{p}_1^{1/2} \hat{\gamma}_{j|\mathcal{F},1}, \dots, \hat{p}_H^{1/2} \hat{\gamma}_{j|\mathcal{F},H})^T$. Then $T_{j|\mathcal{F}}^{\text{SIR}} = n \sum_{h=1}^H \hat{p}_h \hat{\gamma}_{j|\mathcal{F},h}^2$ can be written as $n(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SIR}})^T \hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SIR}}$. Because $\gamma_{j|\mathcal{F},h} = 0$ with $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$ and condition (2.5), $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SIR}}$ has expansion

$$\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SIR}} = \mathbf{L}_{j|\mathcal{F}}^{\text{SIR}} + E_n\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^*\} + o_P(n^{-1/2}),$$

where $(\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^* = (p_1^{1/2} \gamma_{j|\mathcal{F},1}^*, \dots, p_H^{1/2} \gamma_{j|\mathcal{F},H}^*)^T$, and $\gamma_{j|\mathcal{F},h}^*$ is provided in Lemma 1. Define $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{SIR}} = E((\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^* \{(\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^*\}^T)$, and the result of Theorem 3.1 follows directly. \square

Recall the following definitions involved in Theorem 3.2: $\boldsymbol{\nu}_{j|\mathcal{F},h} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} E(\mathbf{X}_{\mathcal{F}} \gamma_{j|\mathcal{F}} | Y \in J_h)$, $\zeta_{j|\mathcal{F},h} = E(\gamma_{j|\mathcal{F}}^2 | Y \in J_h)$, $\boldsymbol{\phi}_{j|\mathcal{F},h} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} \{ \mathbf{U}_{\mathcal{F},h} \gamma_{j|\mathcal{F},h} - E(\mathbf{X}_{\mathcal{F}} \gamma_{j|\mathcal{F}} | Y \in J_h) \}$, and $\mathbf{V}_{\mathcal{F},h} = E(\mathbf{X}_{\mathcal{F}} \mathbf{X}_{\mathcal{F}}^T | Y \in J_h)$.

Lemma 2. Assume the conditions of Theorem 3.2 are satisfied. Under $H_0 : Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$, $j \in \mathcal{F}^c$, the expansions of $\hat{\mathbf{V}}_{\mathcal{F},h}$, $\hat{\boldsymbol{\nu}}_{j|\mathcal{F},h}$, $\hat{\boldsymbol{\phi}}_{j|\mathcal{F},h}$, $\hat{\zeta}_{j|\mathcal{F},h}$ have the form (8.11), where we replace $G(F)$ with $\mathbf{V}_{\mathcal{F},h}$, $\boldsymbol{\nu}_{j|\mathcal{F},h}$, $\boldsymbol{\phi}_{j|\mathcal{F},h}$, or $\zeta_{j|\mathcal{F},h}$, and we replace $G^*(F)$ with $\mathbf{V}_{\mathcal{F},h}^* =$

$$(\mathbf{X}_{\mathcal{F}}\mathbf{X}_{\mathcal{F}}^T R_h - \mathbf{V}_{\mathcal{F},h})/p_h - \mathbf{X}_{\mathcal{F}}(\mathbf{U}_{\mathcal{F},h}^*)^T - (\mathbf{U}_{\mathcal{F},h}^*)\mathbf{X}_{\mathcal{F}}^T, \boldsymbol{\nu}_{j|\mathcal{F},h}^* = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}(x_j\mathbf{X}_{\mathcal{F}} - E(x_j\mathbf{X}_{\mathcal{F}}) - \mathbf{V}_{\mathcal{F}}^*\boldsymbol{\vartheta}_{j|\mathcal{F}} - \mathbf{V}_{\mathcal{F}}\boldsymbol{\vartheta}_{j|\mathcal{F}}^*)/\sigma_{j|\mathcal{F}}, \phi_{j|\mathcal{F},h}^* = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}\mathbf{U}_{\mathcal{F},h}\gamma_{j|\mathcal{F},h}^* - \boldsymbol{\nu}_{j|\mathcal{F},h}^*, \text{ or } \zeta_{j|\mathcal{F},h}^* = \{(-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)\mathbf{V}_{\mathcal{F}\cup j,h}^*(-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)^T + 2(-(\boldsymbol{\vartheta}_{j|\mathcal{F}}^*)^T, 1)\mathbf{V}_{\mathcal{F}\cup j,h}(-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)^T - (\sigma_{j|\mathcal{F}}^2)^*\}/\sigma_{j|\mathcal{F}}^2.$$

PROOF OF LEMMA 2. The expansion of $\hat{\mathbf{V}}_{\mathcal{F},h}$ is parallel to the expansion of $\hat{\mathbf{V}}_h$ in Lemma 1 of Li and Wang (2007). The expansion of $\hat{\boldsymbol{\nu}}_{j|\mathcal{F},h}$ uses the same technique as the expansion of $\hat{\gamma}_{j|\mathcal{F},h}$ in Lemma 1. The expansion of $\hat{\boldsymbol{\phi}}_{j|\mathcal{F},h}$ is obvious by noticing that $\gamma_{j|\mathcal{F},h} = 0$ with $Y \perp\!\!\!\perp x_j|\mathbf{X}_{\mathcal{F}}$ and condition (2.5). For the expansion of $\hat{\zeta}_{j|\mathcal{F},h}$, notice that

$$\sigma_{j|\mathcal{F}}^2 \zeta_{j|\mathcal{F},h} = E((x_j - \boldsymbol{\vartheta}_{j|\mathcal{F}}^T \mathbf{X}_{\mathcal{F}})^2 | Y \in J_h) = (-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)\mathbf{V}_{\mathcal{F}\cup j,h}(-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)^T.$$

We have seen in the proof of Theorem 2.2 that $\zeta_{j|\mathcal{F},h} = 1$ with $Y \perp\!\!\!\perp x_j|\mathbf{X}_{\mathcal{F}}$, conditions (2.5) and (2.6). Taking Frechet derivative on both sides of the listed equation above, we get the desired result. \square

PROOF OF THEOREM 3.2. Let $(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE}})^T = \{(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE1}})^T, (\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE2}})^T\}$ with $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE1}} = \{\hat{p}_1^{1/2}(1 - \hat{\zeta}_{j|\mathcal{F},1} + \hat{\gamma}_{j|\mathcal{F},1}^2), \dots, \hat{p}_H^{1/2}(1 - \hat{\zeta}_{j|\mathcal{F},H} + \hat{\gamma}_{j|\mathcal{F},H}^2)\}^T$ and $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE2}} = (\sqrt{2}\hat{p}_1^{1/2}\hat{\boldsymbol{\phi}}_{j|\mathcal{F},1}^T, \dots, \sqrt{2}\hat{p}_H^{1/2}\hat{\boldsymbol{\phi}}_{j|\mathcal{F},H}^T)^T$. Then $T_{j|\mathcal{F}}^{\text{SAVE}} = n(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE}})^T \hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE}}$. Let $(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE1}})^* = (p_1^{1/2}\zeta_{j|\mathcal{F},1}^*, \dots, p_H^{1/2}\zeta_{j|\mathcal{F},H}^*)^T$, $(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE2}})^* = \{\sqrt{2}p_1^{1/2}(\boldsymbol{\phi}_{j|\mathcal{F},1}^*)^T, \dots, \sqrt{2}p_H^{1/2}(\boldsymbol{\phi}_{j|\mathcal{F},H}^*)^T\}^T$, and $\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^*\}^T = (\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE1}})^*\}^T, \{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE2}})^*\}^T)$. Here $\boldsymbol{\phi}_{j|\mathcal{F},h}^*$ and $\zeta_{j|\mathcal{F},h}^*$ are provided in Lemma 2. With $Y \perp\!\!\!\perp x_j|\mathbf{X}_{\mathcal{F}}$, conditions (2.5) and (2.6), we have $\gamma_{j|\mathcal{F},h} = 0$ and $\zeta_{j|\mathcal{F},h} = 1$. It follows that

$$\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{SAVE}} = \mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}} + E_n\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^*\} + o_P(n^{-1/2}).$$

Let $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{SAVE}} = E((\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^*\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^*\}^T)$. The result of Theorem 3.2 follows directly. \square

Recall the following definitions involved in Theorem 3.3: $\boldsymbol{\iota}_{j|\mathcal{F},h} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} \mathbf{U}_{\mathcal{F},h} \gamma_{j|\mathcal{F},h}$, $\boldsymbol{\iota}_{j|\mathcal{F}} = \sum_{h=1}^H p_h \boldsymbol{\iota}_{j|\mathcal{F},h}$, $\boldsymbol{\varrho}_{j|\mathcal{F}} = \sum_{h=1}^H p_h \gamma_{j|\mathcal{F},h}^2$, and $\kappa_{\mathcal{F}} = \sum_{h=1}^H p_h \mathbf{U}_{\mathcal{F},h}^T \boldsymbol{\Sigma}_{\mathcal{F}}^{-1} \mathbf{U}_{\mathcal{F},h}$. Proof of the following lemma is obvious and omitted.

Lemma 3. *Assume the conditions of Theorem 3.3 are satisfied. Under $H_0 : Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}, j \in \mathcal{F}^c$, the expansion of $\hat{\boldsymbol{\iota}}_{j|\mathcal{F},h}$ has the form (8.11), where we replace $G(F)$ with $\boldsymbol{\iota}_{j|\mathcal{F},h}$, and we replace $G^*(F)$ with $\boldsymbol{\iota}_{j|\mathcal{F},h}^* = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} \mathbf{U}_{\mathcal{F},h} \gamma_{j|\mathcal{F},h}^*$.*

PROOF OF THEOREM 3.3. Let $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR1}} = \{\sqrt{2}\hat{p}_1^{1/2}(1 - \hat{\zeta}_{j|\mathcal{F},1}), \dots, \sqrt{2}\hat{p}_H^{1/2}(1 - \hat{\zeta}_{j|\mathcal{F},H})\}^T$, $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR2}} = \{2\hat{p}_1^{1/2}\hat{\boldsymbol{\nu}}_{j|\mathcal{F},1}^T, \dots, 2\hat{p}_H^{1/2}\hat{\boldsymbol{\nu}}_{j|\mathcal{F},H}^T\}^T$, $\hat{\ell}_{j|\mathcal{F}}^{\text{DR3}} = 2\sum_{h=1}^H \hat{p}_h \hat{\gamma}_{j|\mathcal{F},h}^2$, $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR4}} = 2\sum_{h=1}^H \hat{p}_h \hat{\boldsymbol{\iota}}_{j|\mathcal{F},h}$, and $\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR5}} = \{2(\hat{\kappa}_{\mathcal{F}}\hat{p}_1)^{1/2}\hat{\gamma}_{j|\mathcal{F},1}, \dots, 2(\hat{\kappa}_{\mathcal{F}}\hat{p}_H)^{1/2}\hat{\gamma}_{j|\mathcal{F},H}\}^T$. Then $T_{j|\mathcal{F}}^{\text{DR}} = n(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR}})^T \hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR}}$ with

$$\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR}} = \{(\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR1}})^T, (\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR2}})^T, \hat{\ell}_{j|\mathcal{F}}^{\text{DR3}}, (\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR4}})^T, (\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR5}})^T\}^T.$$

Let $(\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^* = \{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR1}})^*\}^T, \{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR2}})^*\}^T, (\ell_{j|\mathcal{F}}^{\text{DR3}})^*, \{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR4}})^*\}^T, \{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR5}})^*\}^T\}^T$, where $(\mathbf{L}_{j|\mathcal{F}}^{\text{DR1}})^* = \{\sqrt{2}p_1^{1/2}(1 - \zeta_{j|\mathcal{F},1})^*, \dots, \sqrt{2}p_H^{1/2}(1 - \zeta_{j|\mathcal{F},H})^*\}^T$, $(\mathbf{L}_{j|\mathcal{F}}^{\text{DR2}})^* = \{2p_1^{1/2}(\boldsymbol{\nu}_{j|\mathcal{F},1}^*)^T, \dots, 2p_H^{1/2}(\boldsymbol{\nu}_{j|\mathcal{F},H}^*)^T\}^T$, $(\ell_{j|\mathcal{F}}^{\text{DR3}})^* = 0$, $(\mathbf{L}_{j|\mathcal{F}}^{\text{DR4}})^* = 2\sum_{h=1}^H p_h \boldsymbol{\iota}_{j|\mathcal{F},h}^*$, and $(\mathbf{L}_{j|\mathcal{F}}^{\text{DR5}})^* = \{2(\kappa_{\mathcal{F}}p_1)^{1/2}(\gamma_{j|\mathcal{F},1}^*)^T, \dots, 2(\kappa_{\mathcal{F}}p_H)^{1/2}(\gamma_{j|\mathcal{F},H}^*)^T\}^T$. Here $\gamma_{j|\mathcal{F},h}^*$ is provided in Lemma 1, $\boldsymbol{\nu}_{j|\mathcal{F},h}^*$ and $\zeta_{j|\mathcal{F},h}^*$ are provided in Lemma 2, and $\boldsymbol{\iota}_{j|\mathcal{F},h}^*$ is provided in Lemma 3. With $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$, conditions (2.5) and (2.6), we have $\gamma_{j|\mathcal{F},h} = 0$, $\zeta_{j|\mathcal{F},h} = 1$, $\boldsymbol{\nu}_{j|\mathcal{F},h} = \mathbf{0}$. It follows that

$$\hat{\mathbf{L}}_{j|\mathcal{F}}^{\text{DR}} = \mathbf{L}_{j|\mathcal{F}}^{\text{DR}} + E_n\{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^*\} + o_P(n^{-1/2}).$$

Define $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{DR}} = E((\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^*\{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^*\}^T)$, and the result of Theorem 3.3 follows directly. \square

With the expansion forms in the proofs of Theorems 3.1, 3.2 and 3.3, we construct consistent estimators for $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{SIR}}$, $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{SAVE}}$ and $\boldsymbol{\Omega}_{j|\mathcal{F}}^{\text{DR}}$ as follows: $\hat{\boldsymbol{\Omega}}_{j|\mathcal{F}}^{\text{SIR}} = E_n((\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^*\{(\mathbf{L}_{j|\mathcal{F}}^{\text{SIR}})^*\}^T)$,

$\hat{\mathbf{\Omega}}_{j|\mathcal{F}}^{\text{SAVE}} = E_n((\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^* \{(\mathbf{L}_{j|\mathcal{F}}^{\text{SAVE}})^*\}^T)$, and $\hat{\mathbf{\Omega}}_{j|\mathcal{F}}^{\text{DR}} = E_n((\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^* \{(\mathbf{L}_{j|\mathcal{F}}^{\text{DR}})^*\}^T)$. Then the estimated weights $\hat{\omega}_{j|\mathcal{F},k}^{\text{SIR}}$, $\hat{\omega}_{j|\mathcal{F},k}^{\text{SAVE}}$ and $\hat{\omega}_{j|\mathcal{F},k}^{\text{DR}}$ are the k th eigenvalue of $\hat{\mathbf{\Omega}}_{j|\mathcal{F}}^{\text{SIR}}$, $\hat{\mathbf{\Omega}}_{j|\mathcal{F}}^{\text{SAVE}}$ and $\hat{\mathbf{\Omega}}_{j|\mathcal{F}}^{\text{DR}}$ respectively.

8.3. Proofs of Theorems in Section 4

PROOF OF PROPOSITION 4.1. For any $j \in \mathcal{F}^c \cap \mathcal{A}$, we know from Theorem 2.1 and the fact $x_{j|\mathcal{F}} = x_j - E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1} \mathbf{X}_{\mathcal{F}}$ that

$$\begin{aligned} & \sigma_{j|\mathcal{F}}^2 \{ \text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) \} = \text{Var}(x_{j|\mathcal{F}} | Y) \\ &= \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \} \text{Var} \{ E((\mathbf{X}_{\mathcal{F}}^T, x_j) | Y) \} \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \}^T \\ &= \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \} \mathbf{P} \text{Var} \{ E(\mathbf{X} | Y) \} \mathbf{P}^T \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \}^T. \end{aligned} \quad (8.12)$$

Here $\mathbf{P} = (\mathbf{I}_{|\mathcal{F}|+1}, \mathbf{0}_{(|\mathcal{F}|+1) \times (p-|\mathcal{F}|-1)})$, and we assume without loss of generality that the first $|\mathcal{F}| + 1$ elements of \mathbf{X}^T are $(\mathbf{X}_{\mathcal{F}}^T, x_j)$. Since $\mathbf{M}^{\text{SIR}} = \text{Var} \{ E(\mathbf{Z} | Y) \} = \sum_{i=1}^q \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T$ and $\boldsymbol{\beta}_i = \mathbf{\Sigma}^{-1/2} \boldsymbol{\eta}_i$, we have

$$\text{Var} \{ E(\mathbf{X} | Y) \} = \mathbf{\Sigma}^{1/2} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \right) \mathbf{\Sigma}^{1/2} = \mathbf{\Sigma} \left(\sum_{i=1}^q \lambda_i \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \right) \mathbf{\Sigma}. \quad (8.13)$$

Recall that $\mathcal{I} = \{1, \dots, p\}$. Denote $\mathbf{\Sigma}_{\mathcal{F}_1, \mathcal{F}_2} = E(\mathbf{X}_{\mathcal{F}_1}^T \mathbf{X}_{\mathcal{F}_2})$ for any $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{I}$. It follows

$$\begin{aligned} & \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \} \mathbf{P} \mathbf{\Sigma} \boldsymbol{\beta}_i = \{ -E^T(x_j \mathbf{X}_{\mathcal{F}}) \mathbf{\Sigma}_{\mathcal{F}}^{-1}, 1 \} \mathbf{\Sigma}_{\mathcal{F} \cup j, \mathcal{I}} \boldsymbol{\beta}_i \\ &= (\mathbf{\Sigma}_{j, \mathcal{I}} - \mathbf{\Sigma}_{j, \mathcal{F}} \mathbf{\Sigma}_{\mathcal{F}}^{-1} \mathbf{\Sigma}_{\mathcal{F}, \mathcal{I}}) \boldsymbol{\beta}_i = (\mathbf{\Sigma}_{j, \mathcal{F}^c} - \mathbf{\Sigma}_{j, \mathcal{F}} \mathbf{\Sigma}_{\mathcal{F}}^{-1} \mathbf{\Sigma}_{\mathcal{F}, \mathcal{F}^c}) \boldsymbol{\beta}_{i, \mathcal{F}^c}, \end{aligned}$$

where the last equality is true because $(\mathbf{\Sigma}_{j, \mathcal{F}} - \mathbf{\Sigma}_{j, \mathcal{F}} \mathbf{\Sigma}_{\mathcal{F}}^{-1} \mathbf{\Sigma}_{\mathcal{F}, \mathcal{F}}) \boldsymbol{\beta}_{i, \mathcal{F}} = \mathbf{0}$. By the definition of \mathcal{A} in (1.1) and the fact that $\text{Span}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q) = \mathcal{S}_{Y|\mathbf{X}}$, we know $\boldsymbol{\beta}_{i, \mathcal{F}^c \cap \mathcal{A}^c} = \mathbf{0}$.

Thus for $i = 1, \dots, q$,

$$\{-E^T(x_j \mathbf{X}_{\mathcal{F}}) \Sigma_{\mathcal{F}}^{-1}, 1\} \mathbf{P} \Sigma \beta_i = (\Sigma_{j, \mathcal{F}^c \cap \mathcal{A}} - \Sigma_{j, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}}) \beta_{i, \mathcal{F}^c \cap \mathcal{A}}. \quad (8.14)$$

(8.12), (8.13) and (8.14) together imply that

$$\sigma_{j|\mathcal{F}}^2 \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} = \sum_{i=1}^q \lambda_i \{(\Sigma_{j, \mathcal{F}^c \cap \mathcal{A}} - \Sigma_{j, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}}) \beta_{i, \mathcal{F}^c \cap \mathcal{A}}\}^2.$$

By noticing that $\sum_{j \in \mathcal{F}^c \cap \mathcal{A}} \{(\Sigma_{j, \mathcal{F}^c \cap \mathcal{A}} - \Sigma_{j, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}}) \beta_{i, \mathcal{F}^c \cap \mathcal{A}}\}^2 = \beta_{i, \mathcal{F}^c \cap \mathcal{A}}^T (\Sigma_{\mathcal{F}^c \cap \mathcal{A}} - \Sigma_{\mathcal{F}^c \cap \mathcal{A}, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}})^2 \beta_{i, \mathcal{F}^c \cap \mathcal{A}}$, and $\lambda_{\min}(\Sigma_{\mathcal{F}^c \cap \mathcal{A}} - \Sigma_{\mathcal{F}^c \cap \mathcal{A}, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}}) = \lambda_{\max}^{-1}\{(\Sigma_{\mathcal{F}^c \cap \mathcal{A}} - \Sigma_{\mathcal{F}^c \cap \mathcal{A}, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}})^{-1}\} \geq \lambda_{\max}^{-1}(\Sigma^{-1}) = \lambda_{\min}(\Sigma)$, we have

$$\begin{aligned} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \sigma_{j|\mathcal{F}}^2 \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} &\geq |\mathcal{F}^c \cap \mathcal{A}|^{-1} \sum_{j \in \mathcal{F}^c \cap \mathcal{A}} [\sigma_{j|\mathcal{F}}^2 \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\}] \\ &= |\mathcal{F}^c \cap \mathcal{A}|^{-1} \sum_{i=1}^q \lambda_i \beta_{i, \mathcal{F}^c \cap \mathcal{A}}^T (\Sigma_{\mathcal{F}^c \cap \mathcal{A}} - \Sigma_{\mathcal{F}^c \cap \mathcal{A}, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}})^2 \beta_{i, \mathcal{F}^c \cap \mathcal{A}} \\ &\geq |\mathcal{F}^c \cap \mathcal{A}|^{-1} \sum_{i=1}^q \lambda_i \lambda_{\min}^2(\Sigma_{\mathcal{F}^c \cap \mathcal{A}} - \Sigma_{\mathcal{F}^c \cap \mathcal{A}, \mathcal{F}} \Sigma_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}, \mathcal{F}^c \cap \mathcal{A}}) \beta_{i, \mathcal{F}^c \cap \mathcal{A}}^T \beta_{i, \mathcal{F}^c \cap \mathcal{A}} \\ &\geq \lambda_q \lambda_{\min}^2(\Sigma) |\mathcal{F}^c \cap \mathcal{A}|^{-1} \sum_{i=1}^q \beta_{i, \mathcal{F}^c \cap \mathcal{A}}^T \beta_{i, \mathcal{F}^c \cap \mathcal{A}} \geq \lambda_q \lambda_{\min}^2(\Sigma) \beta_{\min}^2. \end{aligned}$$

The proof is then completed by noting that $\max_{j \in \mathcal{F}^c \cap \mathcal{A}} \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} \geq \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \sigma_{j|\mathcal{F}}^2 \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} / \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \sigma_{j|\mathcal{F}}^2$ and $\sigma_{j|\mathcal{F}}^2 \leq \text{Var}(x_j) \leq \lambda_{\max}(\Sigma)$. \square

PROOF OF THEOREM 4.1. For part 1, denote $\Delta = \varsigma n^{-\xi_{\min}} - n^{-1} \bar{c}^{\text{SIR}} > 0$. Because $0 < \bar{c}^{\text{SIR}} < \varsigma n^{1-\xi_{\min}}/2$, we have $\Delta = O_P(n^{-\xi_{\min}})$. When $\mathcal{F}^c \cap \mathcal{A} \neq \emptyset$, $\{\text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}})\} - \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}})\} = O_P(n^{-1/2})$. Note that $0 < \xi_{\min} < 1/2$. Thus as n goes to

infinity, with probability approaching 1,

$$\max_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \left[\left\{ \text{tr}(\mathbf{M}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) \right\} - \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) \right\} \right] < \Delta.$$

Together with (4.3), we know with probability approaching 1,

$$\begin{aligned} & \min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) \right\} > \min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \left\{ \text{tr}(\mathbf{M}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) \right\} \\ & - \max_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \left[\left\{ \text{tr}(\mathbf{M}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) \right\} - \left\{ \text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup \mathcal{A}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) \right\} \right] \\ & > \varsigma n^{-\xi_{\min}} - \Delta = n^{-1} \bar{c}^{\text{SIR}}. \end{aligned}$$

Multiply both sides by n and we get $Pr(\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} T_{j|\mathcal{F}}^{\text{SIR}} > \bar{c}^{\text{SIR}}) \rightarrow 1$.

In part 2, note that $\mathcal{F}^c \cap \mathcal{A} = \emptyset$ implies $\mathcal{A} \subseteq \mathcal{F}$. Then $j \in \mathcal{F}$ implies either $j \in \mathcal{A}$ or $j \in \{\mathcal{F} \setminus \mathcal{A}\}$. If $j \in \mathcal{A}$, then $\{\mathcal{F} \setminus j\}^c \cap \mathcal{A} \neq \emptyset$. As n goes to infinity, $\text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\{\mathcal{F} \setminus j\}}^{\text{SIR}})$ converges to $\text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\{\mathcal{F} \setminus j\}}^{\text{SIR}})$. Condition (4.3) implies that for any $j \in \mathcal{A}$, $T_{j|\{\mathcal{F} \setminus j\}}^{\text{SIR}} > \varsigma n^{1-\xi_{\min}}/2$ with probability 1. If $j \in \{\mathcal{F} \setminus \mathcal{A}\}$, Theorem 3.1 guarantees that $T_{j|\{\mathcal{F} \setminus j\}}^{\text{SIR}}$ converges to a sum of weighted χ^2 , which is $O_p(1)$ and is asymptotically smaller than $\varsigma n^{1-\xi_{\min}}/2$. Thus for $\mathcal{F}^c \cap \mathcal{A} = \emptyset$, $\min_{j \in \mathcal{F}} T_{j|\{\mathcal{F} \setminus j\}}^{\text{SIR}} = O_p(1) < C n^{1-\xi_{\min}}$ with $\xi_{\min} < 1$ and $C > 0$. It follows that $Pr(\max_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} = \emptyset} \min_{j \in \mathcal{F}} T_{j|\{\mathcal{F} \setminus j\}}^{\text{SIR}} < \underline{c}^{\text{SIR}}) \rightarrow 1$ if we set $\underline{c}^{\text{SIR}} > C n^{1-\xi_{\min}}$. \square

For stepwise SAVE, we replace condition (4.3) with

$$\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \left\{ \text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SAVE}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SAVE}}) \right\} > \varsigma n^{-\xi_{\min}}. \quad (8.15)$$

For stepwise directional regression, we replace condition (4.3) with

$$\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{F}^c \cap \mathcal{A}} \{\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{DR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{DR}})\} > \varsigma n^{-\xi_{\min}}. \quad (8.16)$$

Conditions (8.15) and (8.16) are parallel to condition (4.3), and will guarantee the selection consistency of stepwise SAVE and stepwise directional regression respectively.

Before proceeding to the proof of Theorem 5.1, we present the following useful lemmas. For $\mathbf{a} = (a_1, \dots, a_p)^T \in \mathbb{R}^p$, let $\|\mathbf{a}\|_{\infty} = \max_{1 \leq i \leq p} |a_i|$, $\|\mathbf{a}\|_1 = \sum_{i=1}^p |a_i|$, and $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^p a_i^2}$. For $\mathbf{A} = \{a_{ij}\} \in \mathbb{R}^{p \times p}$, let $\|\mathbf{A}\|_{\infty} = \max_{1 \leq i, j \leq p} |a_{ij}|$.

Lemma 4. *For $h = 1, \dots, H$, let $W_h = p_h^{-1/2} (I(Y \in J_h) - p_h)$ and $\boldsymbol{\alpha}_h = \boldsymbol{\Sigma}^{-1} E(\mathbf{X}W_h)$. Then $\text{tr}(\mathbf{M}^{\text{SIR}}) = (H - 1) - \sum_{h=1}^H E(W_h - \mathbf{X}^T \boldsymbol{\alpha}_h)^2$.*

PROOF OF LEMMA 4. We notice that for $h = 1, \dots, H$,

$$\begin{aligned} E(W_h - \mathbf{X}^T \boldsymbol{\alpha}_h)^2 &= E(W_h^2) + E(\boldsymbol{\alpha}_h^T \mathbf{X} \mathbf{X}^T \boldsymbol{\alpha}_h) - 2E(W_h \mathbf{X}^T \boldsymbol{\alpha}_h) \\ &= E(W_h^2) - \boldsymbol{\alpha}_h^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_h = (1 - p_h) - p_h^{-1} E\{\mathbf{Z}^T I(Y \in J_h)\} E\{\mathbf{Z} I(Y \in J_h)\}. \end{aligned}$$

Add over h and we get the desired result. \square

Lemma 5. *Let $\mathbf{W}^{\text{SIR}} = \text{Var}\{E(\mathbf{X}|Y)\}$ and let $\hat{\mathbf{W}}^{\text{SIR}}$ be its corresponding sample estimator. Denote $p_0 = \min_{1 \leq h \leq H} p_h$, $\tau_0 = 1.25 \exp\{1 + (10\tau_{\min})^{-1/2} \tau_{\max}^{1/2}\}$, $D_1 = 2 + 8\tau_0^2$, $D_2 = (10\tau_{\min})^{1/2} D_1$, $D_3 = 2H\tau_{\max}^{1/2} p_0^{-1} D_2$ and $D_4 = D_3 + H p_0^{-1} D_2^2 + (2 + 12.5e^2)^2$. Then under the same conditions of Theorem 5.1, we have $\max_{1 \leq i, j \leq p} |\hat{\mathbf{W}}_{ij}^{\text{SIR}} - \mathbf{W}_{ij}^{\text{SIR}}| \leq D_4 \sqrt{\log p/n}$ with probability tending to 1.*

PROOF OF LEMMA 5. Let $\mathbf{u}_h = E\{\mathbf{X}I(Y \in J_h)\}$ and let $\hat{\mathbf{u}}_h = E_n\{\mathbf{X}I(Y \in J_h)\}$ be its

sample estimator. By the definition of \mathbf{W}^{SIR} and $\hat{\mathbf{W}}^{\text{SIR}}$, we have

$$\hat{\mathbf{W}}^{\text{SIR}} - \mathbf{W}^{\text{SIR}} = \sum_{h=1}^H \{\hat{p}_h^{-1} \hat{\mathbf{u}}_h \hat{\mathbf{u}}_h^T - p_h^{-1} \mathbf{u}_h \mathbf{u}_h^T\} - E_n(\mathbf{X}) E_n(\mathbf{X}^T).$$

Denote $\mathbf{W}^{(1)} = \sum_{h=1}^H p_h^{-1} (\hat{\mathbf{u}}_h - \mathbf{u}_h) \mathbf{u}_h^T$, $\mathbf{W}^{(2)} = \sum_{h=1}^H p_h^{-1} \mathbf{u}_h (\hat{\mathbf{u}}_h - \mathbf{u}_h)^T$, $\mathbf{W}^{(3)} = \sum_{h=1}^H p_h^{-1} (\hat{\mathbf{u}}_h - \mathbf{u}_h) (\hat{\mathbf{u}}_h - \mathbf{u}_h)^T$, $\mathbf{W}^{(4)} = \sum_{h=1}^H (\hat{p}_h p_h)^{-1} (\hat{p}_h - p_h) \hat{\mathbf{u}}_h \hat{\mathbf{u}}_h^T$, and $\mathbf{W}^{(5)} = E_n(\mathbf{X}) E_n(\mathbf{X}^T)$. Then we have

$$\hat{\mathbf{W}}^{\text{SIR}} - \mathbf{W}^{\text{SIR}} = \mathbf{W}^{(1)} + \mathbf{W}^{(2)} + \mathbf{W}^{(3)} + \mathbf{W}^{(4)} - \mathbf{W}^{(5)}. \quad (8.17)$$

Since $\mathbf{X} = (x_1, \dots, x_p)^T$ is normal, condition (C2) implies that $E\{\exp(tx_i^2)\} \leq 1.25$ for any t such that $0 \leq t \leq (10\tau_{\min})^{-1}$. Inequality $\exp(s) \leq \exp(s^2 + 1)$ implies that $E\{\exp(t|x_i|)\} \leq 1.25e$ as long as $0 \leq t \leq (10\tau_{\min})^{-1}$. Note that $|x_i I(Y \in J_h)| \leq |x_i|$ and $|\mathbf{u}_{h,i}| = |E\{x_i I(Y \in J_h)\}| \leq E(x_i^2)^{1/2} \leq \tau_{\max}^{1/2}$. We have $E\{\exp(t|x_i I(Y \in J_h) - \mathbf{u}_{h,i}|)\} \leq E\exp(t|x_i|) \exp\{(10\tau_{\min})^{-1/2} \tau_{\max}^{1/2}\} \leq 1.25 \exp\{1 + (10\tau_{\min})^{-1/2} \tau_{\max}^{1/2}\}$ for $0 \leq t \leq (10\tau_{\min})^{-1/2}$. Let $\epsilon = (10\tau_{\min})^{-1/2} \sqrt{\log p/n}$. Following similar arguments in the proof of Theorem 1 in Cai et al. (2011), we have

$$\begin{aligned} & Pr(\|\hat{\mathbf{u}}_h - \mathbf{u}_h\|_{\infty} \geq D_2 \sqrt{\log p/n}) \\ & \leq 2p \exp(-D_1 \log p) [E \exp\{\epsilon(x_i I(Y \in J_h) - \mathbf{u}_{h,i})\}]^n \\ & \leq 2p \exp\{-D_1 \log p + n\epsilon^2 E(x_i I(Y \in J_h) - \mathbf{u}_{h,i})^2 \exp(\epsilon|x_i I(Y \in J_h) - \mathbf{u}_{h,i}|)\} \\ & \leq 2p \exp\{-D_1 \log p + 8\tau_0^2 \log p\} = 2p \exp\{-2 \log p\}. \end{aligned}$$

Thus we have

$$Pr(\|\hat{\mathbf{U}}_h - \mathbf{U}_h\|_\infty \geq D_2 \sqrt{\log p/n}) \leq 2p^{-1}. \quad (8.18)$$

Moreover, it is easy to see that

$$\|\mathbf{U}_h\|_\infty \leq \tau_{\max}^{1/2} p_h^{-1} \leq \tau_{\max}^{1/2} p_0^{-1}. \quad (8.19)$$

Combining (8.18) and (8.19), we see with probability tending to 1,

$$\|\mathbf{W}^{(1)} + \mathbf{W}^{(2)}\|_\infty \leq D_3 \sqrt{\log p/n}. \quad (8.20)$$

By (8.18) and condition (C4), we see with probability tending to 1,

$$\|\mathbf{W}^{(3)}\|_\infty \leq H p_0^{-1} D_2^2 \sqrt{\log p/n}. \quad (8.21)$$

Similar to (8.18), we can also show that $Pr(\|E_n(\mathbf{X})\|_\infty \geq \{2 + 8(1.25e)^2\} \sqrt{\log p/n}) \leq 2p^{-1}$. Under condition (C3), we have with probability tending to 1,

$$\|\mathbf{W}^{(5)}\|_\infty \leq (2 + 12.5e^2)^2 \sqrt{\log p/n}. \quad (8.22)$$

Because $\hat{p}_h - p_h = O_P(n^{-1/2})$, we know that $\|\mathbf{W}^{(4)}\|_\infty = O_P(n^{-1/2})$. Together with (8.17), (8.20), (8.21), and (8.22), we get the desired result. \square

Lemma 6. *Assume conditions (C1) and (C2) hold. Then there exists $D_5 > 0$ such that $\max_{1 \leq i, j \leq p} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \leq D_5 \sqrt{\log p/n}$ with probability tending to 1.*

The proof of this Lemma is available in Cai et al. (2011) and thus omitted.

Lemma 7. *Let $\boldsymbol{\delta}_{j|\mathcal{F}} = (-\boldsymbol{\vartheta}_{j|\mathcal{F}}^T, 1)^T$ and $\hat{\boldsymbol{\delta}}_{j|\mathcal{F}} = (-\hat{\boldsymbol{\vartheta}}_{j|\mathcal{F}}^T, 1)^T$. Define $D_6 = 1 + 16\tau_{\min}^{-2}(\tau_{\max})^2$ and $D_0 = D_4D_6 + (4\tau_{\min}^{-2}\tau_{\max}^2 + \tau_{\min}^{-1}\tau_{\max})D_5D_6$. Assume the same conditions of Theorem 5.1 hold. Suppose $|\mathcal{F}| = O(n^{\xi_0 + \xi_{\min}})$. Then with probability tending to 1, we have*

$$|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \hat{\mathbf{W}}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}| \leq D_0 |\mathcal{F}| \sqrt{\log p/n}.$$

PROOF OF LEMMA 7. For any $|\mathcal{F}| = O(n^{\xi_0 + \xi_{\min}})$, we know from Lemm 1 in Wang (2009) that $2^{-1}\tau_{\min} < \lambda_{\min}(\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}) < \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}) < 2\tau_{\max}$. Moreover, $\lambda_{\max}(\mathbf{W}_{\mathcal{F}}^{\text{SIR}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{F}}) < \tau_{\max}$ and $\lambda_{\max}(\hat{\mathbf{W}}_{\mathcal{F}}^{\text{SIR}}) \leq \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}) < 2\tau_{\max}$. It follows that

$$\begin{aligned} \|\boldsymbol{\delta}_{j|\mathcal{F}}\|_2^2 &= 1 + E(x_j \mathbf{X}_{\mathcal{F}}^T) \boldsymbol{\Sigma}_{\mathcal{F}}^{-2} E(x_j \mathbf{X}_{\mathcal{F}}) \leq 1 + \tau_{\min}^{-2} \|E(x_j \mathbf{X}_{\mathcal{F}})\|_2^2 \\ &\leq 1 + \tau_{\min}^{-2} (\tau_{\max})^2, \end{aligned} \quad (8.23)$$

and

$$\begin{aligned} \|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}\|_2^2 &= 1 + E_n(x_j \mathbf{X}_{\mathcal{F}}^T) \hat{\boldsymbol{\Sigma}}_{\mathcal{F}}^{-2} E_n(x_j \mathbf{X}_{\mathcal{F}}) \leq 1 + 4\tau_{\min}^{-2} \|E_n(x_j \mathbf{X}_{\mathcal{F}})\|_2^2 \\ &\leq 1 + 16\tau_{\min}^{-2} (\tau_{\max})^2. \end{aligned} \quad (8.24)$$

By triangular inequality, we have

$$\begin{aligned} |\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \hat{\mathbf{W}}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}| &\leq |\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \hat{\mathbf{W}}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}| \\ &\quad + |\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}| \end{aligned} \quad (8.25)$$

We bound the two terms of (8.25) respectively. Invoking Lemma 5 and (8.24), we have

$$\begin{aligned} |\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \hat{\mathbf{W}}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}| &\leq \|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}\|_{\ell_2}^2 |\mathcal{F}| \max_{1 \leq i, j \leq p} |\hat{\mathbf{W}}_{ij}^{\text{SIR}} - \mathbf{W}_{ij}^{\text{SIR}}| \\ &\leq D_4 D_6 |\mathcal{F}| \sqrt{\log p/n}, \end{aligned} \quad (8.26)$$

and

$$\begin{aligned}
|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}| &= |(\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T + \boldsymbol{\delta}_{j|\mathcal{F}}^T)^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} (\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T - \boldsymbol{\delta}_{j|\mathcal{F}}^T)| \\
&\leq \tau_{\max} \|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T + \boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2 \|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T - \boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2.
\end{aligned} \tag{8.27}$$

By (8.23) and (8.24), we have

$$\|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T + \boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2 \leq (\|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T\|_2 + \|\boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2) \leq D_6. \tag{8.28}$$

Invoking Lemma 5 and 6, we can derive that

$$\begin{aligned}
\|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T - \boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2 &= \|\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}^{-1} E_n(x_j \mathbf{X}_{\mathcal{F}}) - \boldsymbol{\Sigma}_{\mathcal{F}}^{-1} E(x_j \mathbf{X}_{\mathcal{F}})\|_2 \\
&\leq \|\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}^{-1} E_n(x_j \mathbf{X}_{\mathcal{F}}) - \boldsymbol{\Sigma}_{\mathcal{F}}^{-1} E_n(x_j \mathbf{X}_{\mathcal{F}})\|_2 + \|\boldsymbol{\Sigma}_{\mathcal{F}}^{-1} E_n(x_j \mathbf{X}_{\mathcal{F}}) - \boldsymbol{\Sigma}_{\mathcal{F}}^{-1} E(x_j \mathbf{X}_{\mathcal{F}})\|_2 \\
&\leq \lambda_{\max}^{1/2} \{(\hat{\boldsymbol{\Sigma}}_{\mathcal{F}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{F}}^{-1})^2\} \|E_n(x_j \mathbf{X}_{\mathcal{F}})\|_2 + \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}) |\mathcal{F}|^{1/2} \|E_n(x_j \mathbf{X}_{\mathcal{F}}) - E(x_j \mathbf{X}_{\mathcal{F}})\|_{\infty} \\
&\leq 2\tau_{\min}^{-2} |\mathcal{F}| D_5 \sqrt{\log p/n} \cdot 2\tau_{\max} + \tau_{\min}^{-1} |\mathcal{F}|^{1/2} D_5 \sqrt{\log p/n}.
\end{aligned}$$

It follows that

$$\|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T - \boldsymbol{\delta}_{j|\mathcal{F}}^T\|_2 \leq (4\tau_{\min}^{-2} \tau_{\max} + \tau_{\min}^{-1}) D_5 \sqrt{\log p/n}. \tag{8.29}$$

(8.28) and (8.29) together suggest that

$$|\hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}} - \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}| \leq (4\tau_{\min}^{-2} \tau_{\max}^2 + \tau_{\min}^{-1} \tau_{\max}) D_5 D_6 |\mathcal{F}| \sqrt{\log p/n}. \tag{8.30}$$

Plug (8.26), (8.27) and (8.30) into (8.25), and we get the desired result. \square

PROOF OF THEOREM 5.1. Let $C_0 = 2H\varsigma^{-1}$. We first state the outline of the proof as follows. Notice that $|\mathcal{A}| \leq \varpi n^{\xi_0}$ from condition (C4). To include $|\mathcal{A}|$ relevant predictors in the FTP algorithm within $[C_0\varpi n^{\xi_0+\xi_{\min}}]$ steps, all we need to show is that within $[C_0n^{\xi_{\min}}]$ steps, at least one new significant variable will be selected by the FTP algorithm, conditional on those already included. A complete proof would entail $|\mathcal{A}|$ stages, with each stage focusing on the i th block of $[C_0n^{\xi_{\min}}]$ steps in the FTP algorithm, $i = 1, \dots, |\mathcal{A}|$. Without loss of generality, we focus on the first block of $[C_0n^{\xi_{\min}}]$ steps in the FTP algorithm, and show that at least one significant variable will be included.

Assume no relevant predictors have been selected in the first k steps, and we evaluate what happens at the $k + 1$ th step. Define

$$\Omega(k) = \text{tr} \left(\hat{\mathbf{M}}_{\mathcal{S}^{(k)}}^{\text{SIR}} \right) - \text{tr} \left(\hat{\mathbf{M}}_{\mathcal{S}^{(k-1)}}^{\text{SIR}} \right), k = 1, 2, \dots, [C_0n^{\xi_{\min}}].$$

From this definition, we have $\sum_{k=1}^{[C_0n^{\xi_{\min}}]} \Omega(k) = \text{tr} \left(\hat{\mathbf{M}}_{\mathcal{S}^{([C_0n^{\xi_{\min}}]+1)}}^{\text{SIR}} \right) - \text{tr} \left(\hat{\mathbf{M}}_{\mathcal{S}^{(0)}}^{\text{SIR}} \right)$. Because $\mathcal{S}^{(0)} = \emptyset$, it follows from Lemma 4 that

$$\sum_{k=1}^{[C_0n^{\xi_{\min}}]} \Omega(k) = \text{tr} \left(\hat{\mathbf{M}}_{\mathcal{S}^{([C_0n^{\xi_{\min}}])}}^{\text{SIR}} \right) \leq H - 1. \quad (8.31)$$

We will see later that

$$\Omega(k) \geq \varsigma n^{-\xi_{\min}}/2, \text{ if } a_k \notin \mathcal{A}, \quad k = 1, 2, \dots, [C_0n^{\xi_{\min}}], \quad (8.32)$$

which implies

$$\sum_{k=1}^{\lfloor C_0 n^{\xi_{\min}} \rfloor} \Omega(k) \geq H \text{ if } a_k \notin \mathcal{A}, \quad k = 1, 2, \dots, \lfloor C_0 n^{\xi_{\min}} \rfloor. \quad (8.33)$$

Together, (8.31) and (8.33) imply that there must exist $a_k \in \mathcal{A}$ for some k such that $1 \leq k \leq \lfloor C_0 n^{\xi_{\min}} \rfloor$.

It remains to prove (8.32). By Theorem 3.1, we can derive that $\text{tr}(\mathbf{M}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\text{SIR}}) = \sum_{h=1}^H p_h \gamma_{j|\mathcal{F},h}^2 = \sigma_{j|\mathcal{F}}^{-2} \boldsymbol{\delta}_{j|\mathcal{F}}^T \mathbf{W}_{\mathcal{F} \cup j}^{\text{SIR}} \boldsymbol{\delta}_{j|\mathcal{F}}$ for any \mathcal{F} such that $|\mathcal{F}| < n$. In the sample level, we then have $\text{tr}(\hat{\mathbf{M}}_{\mathcal{F} \cup j}^{\text{SIR}}) - \text{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\text{SIR}}) = \hat{\sigma}_{j|\mathcal{F}}^{-2} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}^T \hat{\mathbf{W}}_{\mathcal{F} \cup j}^{\text{SIR}} \hat{\boldsymbol{\delta}}_{j|\mathcal{F}}$. From the proof of Lemma 3 in Jiang and Liu (2013) and Lemma 6, we know that $\hat{\sigma}_{j|\mathcal{F}}^2 - \sigma_{j|\mathcal{F}}^2 = O(|\mathcal{F}| \sqrt{\log p/n})$. Then under condition (C3), we see that $\hat{\sigma}_{j|\mathcal{F}}^{-2} \geq \sigma_{j|\mathcal{F}}^{-2}/2$ provided that $|\mathcal{F}| = O(n^{\xi_0 + \xi_{\min}})$. Note that $|\mathcal{S}^{(k)}| \leq C_0 \varpi n^{\xi_0 + \xi_{\min}}$. Then by Lemma 7 and condition (C3), we can get

$$\begin{aligned} \Omega(k) &\geq 2^{-1} \sigma_{a_{k+1}|\mathcal{S}^{(k)}}^{-2} (\boldsymbol{\delta}_{a_{k+1}|\mathcal{S}^{(k)}}^T \mathbf{W}_{\mathcal{S}^{(k+1)}}^{\text{SIR}} \boldsymbol{\delta}_{a_{k+1}|\mathcal{S}^{(k)}} - D_0 |\mathcal{S}^{(k+1)}| \sqrt{\log p/n}) \\ &\geq (\varsigma n^{-\xi_{\min}}/2 - 2^{-1} \sigma_{a_{k+1}}^{-2} D_0 \cdot (C_0 \varpi) n^{(\xi_0 + \xi_{\min})} \cdot \varpi^{1/2} n^{\xi/2} n^{-1/2}) \rightarrow \varsigma n^{-\xi_{\min}}/2, \end{aligned}$$

if $a_k \notin \mathcal{A}$, $k = 1, 2, \dots, \lfloor C_0 n^{\xi_{\min}} \rfloor$. The proof is completed. \square

PROOF OF THEOREM 5.2. Define $k_{\min} = \min_{1 \leq k \leq n} \{k : \mathcal{A} \subset \mathcal{S}^{(k)}\}$. Theorem 5.1 guarantees that $k_{\min} \leq 2H\varsigma^{-1} \varpi n^{\xi_0 + \xi_{\min}}$. Following the proof of Theorem 2 in Wang (2009), it's easy to prove that $Pr\left(\min_{0 \leq k < k_{\min}} \{\text{BIC}(\mathcal{S}^{(k)}) - \text{BIC}(\mathcal{S}^{(k+1)})\} > 0\right) \rightarrow 1$, and the details are omitted. \square

References

- Bentler, P. M. and Xie, J. (2000). Corrections to test statistics in principal Hessian direction. *Statistics and Probability Letters* **47** 381–389.
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384.
- Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.
- Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics* **35** 2313–2351.
- Chang, J., Tang, C. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* forthcoming.
- Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika* **95** 759–771.
- Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38** 3696–3723.
- Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley, New York.
- Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics* **32** 1062–1092.
- Cook, R. D. and Li, L. (2009). Dimension reduction in regression with exponential family predictors. *Journal of Computational and Graphical Statistics* **18** 774–791.

- Cook, R. D. and Nachtshim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association* **89** 592-599.
- Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association* **98** 340-351.
- Cook, R. D. and Weisberg, S. (1991). Discussion of “sliced inverse regression for dimension reduction”. *Journal of the American Statistical Association* **86** 328-332.
- Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583-3593.
- Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97** 279-294.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348-1360.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70** 849-911.
- Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer, New York.
- Field, C. (1993). Tail Areas of Linear Combinations of Chi-Squares and Noncentral Chi-Squares. *Journal of Statistical Computation and Simulation* **45** 243-248.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J. and Caligiuri, M. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** 531-536.

- He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41** 342–369.
- Jiang, B. and Liu, J. S. (2013). Sliced inverse regression with variable selection and interaction detection. *Manuscript*.
- Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics* **37** 1272–1298.
- Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102** 997–1008.
- Li, G., Peng, H., Zhang, J. and Zhu, L.-X. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40** 1846–1877.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86** 316–327.
- Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613.
- Li, L., Cook, R.D. and Nachtsheim, C.J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society, Series B* **67** 285–299.
- Li, L. and Yin, X. (2008). Sliced inverse regression with regularization. *Biometrics* **64** 124–131.
- Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association* **107** 1129–1139.
- Lin, L., Sun, J. and Zhu, L. X. (2013). Nonparametric feature screening. *Computational Statistics and Data Analysis*, **67**, 162 – 174.

- Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92** 242–247.
- Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94** 285–296.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288.
- Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104** 1512–1524.
- Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis* **99** 1733–1757.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68** 49–67.
- Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.
- Zhong, W., Zhang, T., Zhu, M. and Liu, J. S. (2012). Correlation pursuit: forward step-wise variable selection for index models. *Journal of Royal Statistical Society: Series B* **74** 849–870.
- Zhou, J. and He, X. (2008). Dimension reduction on constrained canonical correlation and variable filtering. *The Annals of Statistics* **36** 2313–2351.

Zhu, L. P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of American Statistical Association* **106** 1464–1475.